DOCUMENT RESUME

ED 132 177 95 TM 005 892

AUTHOR	Coles, Gary J.; And Others
TITLE	Impact of Educational Innovation on Student Performance: Project Methods and Findings for Three Cohorts. Project LONGSTEP Final Report: Volume I.
INSTITUTION	American Institutes for Research in the Behavioral Sciences, Palo Alto, Calif.
SPONS AGENCY	Office of Education (DHEW), Washington, D.C.
REPORT NO	AIR-21400-4/76-FR-I
PUB DATE	Apr 76
CONTRACT	OEC-0-70-4789
NOTE	400p.; For related documents, see TM 005 891-896 and TM 005 920-922
EDRS PRICE	MF-$0.83 HC-$20.75 Plus Postage.
DESCRIPTORS	*Academic Achievement; Achievement Gains; Achievement Tests; Classroom Observation Techniques; Educational Experience; *Educational Innovation; Elementary School Students; Elementary Secondary Education; Language Arts; *Longitudinal Studies; Mathematics; Questionnaires; Reading Achievement; *Research Methodology; Socioeconomic Status; Standardized Tests; Statistical Analysis; Student Attitudes; Student Characteristics; Teacher Characteristics
IDENTIFIERS	Longitudinal Study of Educational Practices; *Project LONGSTEP

ABSTRACT

The general emphasis of Project LONGSTEP was on the identification of changes in student achievement that occur as a result of intensive educational innovation. This report focuses on three groups of students: those who started out as first-graders, fourth-graders, and sixth-graders in 1970-71. Its primary purpose was to analyze overall differences in achievement growth among educational approaches; an educational growth model in which achievement was related to innovative emphasis, number of minutes of instruction per day, pretest, socioeconomic status, and teaching qualifications; a posttest achievement score statistically adjusted for pretest and socioeconomic differences; groups of students with similar educational experiences who, on the average, performed either much better or much worse than was expected from their pretest and socioeconomic status; and students who demonstrated particularly large achievement gains during two consecutive school years. Even though overall project findings showed that dramatic school effects were not associated with intensity of educational innovation, different educational approaches did produce meaningful and important differences in achievement, especially in the early elementary grades. Unquestionably large gains in reading, language, and arithmetic skills, over and above those expected on the basis of pretest and socioeconomic status, were found to occur in some of the participating schools. "Data Collection Instruments and Guidelines" developed for Project LONGSTEP referenced in Vol. I, Chapter II, Section C, will be accessioned TM 005 987 in RIEMAY77. (RC)

Documents acquired by ERIC include many informal unpublished materials not available from other sources. ERIC makes every effort to obtain the best copy available. Nevertheless, items of marginal reproducibility are often encountered and this affects the quality of the microfiche and hardcopy reproductions ERIC makes available via the ERIC Document Reproduction Service (EDRS). EDRS is not responsible for the quality of the original document. Reproductions supplied by EDRS are the best that can be made from the original.

ED132177

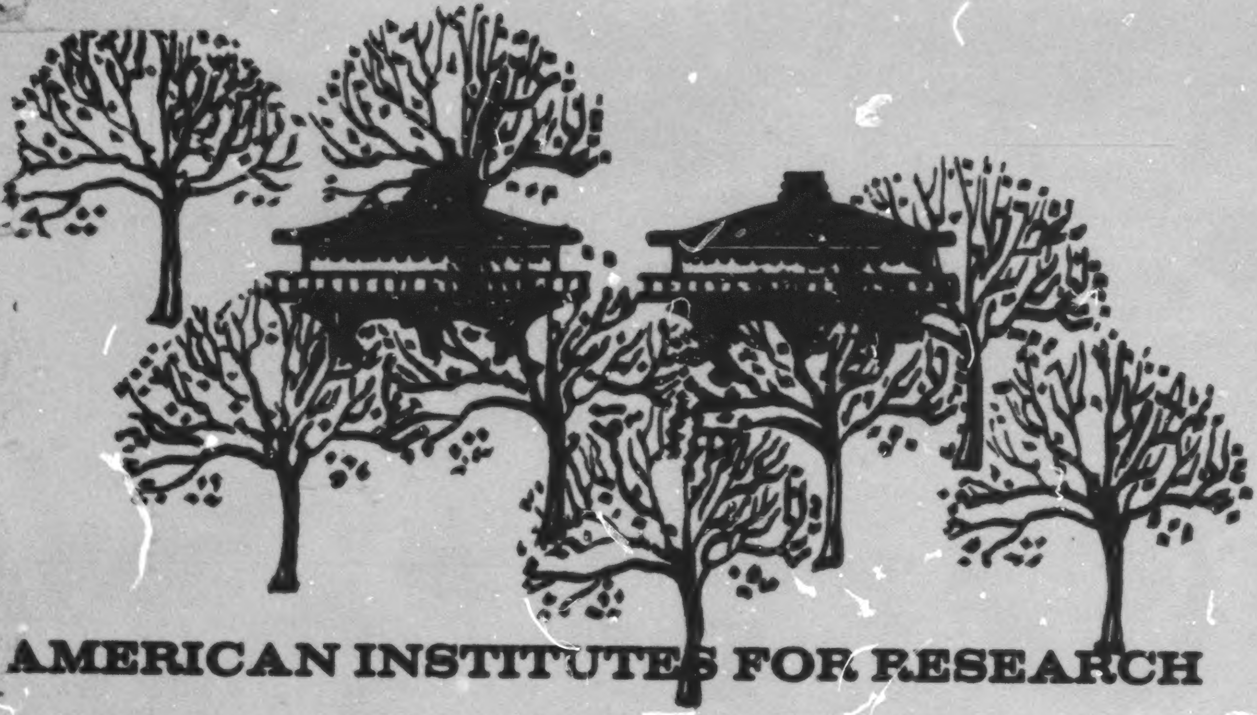
IMPACT OF EDUCATIONAL INNOVATION ON STUDENT PERFORMANCE:

PROJECT METHODS AND FINDINGS FOR THREE COHORTS

Project LONGSTEP Final Report

Volume I

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY



TM005 892

AMERICAN INSTITUTES FOR RESEARCH
Palo Alto, California

AIR—21400—4/76—FR-I

IMPACT OF EDUCATIONAL INNOVATION ON STUDENT PERFORMANCE:

PROJECT METHODS AND FINDINGS FOR THREE COHORTS

Gary J. Coles
Albert B. Chalupsky
Bruce E. Everett
Marion F. Shaycoft
Barbara J. Rodabaugh
Malcolm N. Danoff

Project LONGSTEP Final Report: Volume I

Prepared for

Office of Planning, Budgeting and Evaluation
Office of Education
Washington, D. C. 20202

American Institutes for Research
Palo Alto, California 94302

April 1976

The research reported herein was performed pursuant to Contract No. OEC-0-70-4789 with the U. S. Office of Education, U. S. Department of Health, Education, and Welfare. Contractors undertaking such projects under Government sponsorship are encouraged to express freely their professional judgment in the conduct of the project. Points of view or opinions stated do not, therefore, necessarily represent official Office of Education position or policy.

PREFACE

This document is Volume I of the final report for Project LONGSTEP, the Longitudinal Study of Educational Practices. Together with the separately-bound appendices, it represents the first formal documentation of the project methods and findings. A number of additional topics will be presented in future reports, and an overall summary of the findings will be prepared at the conclusion of the project.

It should be stressed that the present volume, by design, is a technical presentation of the methods and the findings of what is admittedly a very extensive and complex study. Much of the report, therefore, may be of interest primarily to those engaged in educational research and evaluation. For the less technically-inclined reader, a brief overview of project methods and a general summary of our findings and conclusions are contained in the section entitled "Executive Summary."

Any effort of Project LONGSTEP's scope depends on the cooperation of many people and many organizations. We are indebted first of all to the following school districts whose completely voluntary participation in the study was, without doubt, an expression of sincere interest in improving education:

Cashton Public Schools
Cashton, Wisconsin

Duluth Public Schools,
Independent School District No. 709
Duluth, Minnesota

Nova Schools
Fort Lauderdale, Florida

La Farge Area Schools
La Farge, Wisconsin

Laguna Beach Unified School District
Laguna Beach, California

Milton Area School District
Milton, Pennsylvania

Newport-Mesa Unified School District
Newport Beach, California

Norwalk-Ontario School District
Ontario, Wisconsin

Paducah Independent School District
Paducah, Kentucky

Granite School District
Salt Lake City, Utah

San Angelo Public Schools
San Angelo, Texas

Seattle Public Schools
Seattle, Washington

Westby Area Schools
Westby, Wisconsin

To the superintendents, principals, teachers and students in these districts, we extend our sincere appreciation.

Throughout the course of this project we enlisted the help of numerous consultants for their advice and suggestions. In particular, we would like to acknowledge the suggestions from Paul Horst, Robert Gagné, Ralph Tyler, Henry Dyer, Maurice Tatsuoka and Herbert Kiesling during the formative years of Project LONGSTEP. Herbert Walberg and Robert Linn served as reviewers of the draft of this report as did Daryl Nichols, Vincent Campbell and John Bowers of AIR. The report has benefited greatly from their comments; however, they bear no responsibility for any of its defects.

Although they are not listed as authors, a number of AIR staff members contributed significantly to this report and merit special recognition:

> Carolyn L. Davis; for her enthusiastic and tireless efforts in the administrative and clerical areas. The quality of this final report is due, in very large measure, to her careful attention and to her amazing tolerance for numerous "final" changes.
>
> Doris A. Alfred; for her processing, checking, correcting and integrating of an exceedingly large and complicated data base.
>
> Standish W. Sibley; for his work in developing and carefully checking many of the tables shown in the appendices to this volume, in organizing and processing raw data forms, and in critically proofreading the text and tables found throughout this report.
>
> Edwin M. Carr; for his preparation of a number of diagrams and his supervision of the many quality control operations vital to the final production of this report.

During the planning and field implementation years, Project LONGSTEP profited from the ideas and the efforts of many AIR staff members. Malcolm N. Danoff was the Associate Project Director during these years and made significant contributions to the development of data collection instruments and field data collection procedures. David Beswick played a major part in formulating the early project design. The statistical analyses for the interim progress reports benefited greatly from the advice of A. Oscar H. Roberts, who also served as a site visitor. In addition, he was director of a subtask to determine the dependability of our questionnaire data and was senior author of the memorandum report covering this effort. Sarah J. Roberts was a coauthor of the report and was assisted by Laurie Harrison and John Dervin. Mrs. Roberts also served as a field data collector. Another substudy, involving the analysis of alternative classroom observation techniques, was directed by Robert A. Weisgerber, who also was a member of our on-site data collection team. Patrick McDonnell served as director of our initial computer activities. Carol L. Schalla, in addition to her duties as a site visitor, was responsible for organizing the teacher and treatment data and for developing the crossreference codes which permitted treatment and teacher data to be related to students.

Contributing to the very early project design and instrument development activities were Rex C. Fortune and William L. Raley, while Peggie L. Campeau and G. Kasten Tallmadge assisted greatly in the early identification of educational programs. James W. Shearer assisted in questionnaire revision and test coordination, in addition to serving as a site visitor. Another site visitor, Otto t'Hooft contributed to many of the study's early data collection and logistics efforts. Our appreciation is also extended to William M. Shanner for his advice on test selection and attitude measurement. Jack J. Crawford also consulted on the project during its early years.

Paul Whiting, Jr. and Larry Seiders of Comsis Corporation, Mountain View, California, adapted existing file maintenance and crossreferencing computer programs so that they could be used with the complex multilevel data gathered. Computer software developed for Project TALENT was also utilized, as were analysis programs contained in the Statistical Package for the Social Sciences (SPSS) and the UCLA Biomedical Computer Programs (BMD). Almost all of the data processing and analysis involved the computer facilities of Optimum Systems, Incorporated, Santa Clara, California.

The success of Project LONGSTEP's logistics activities was the result of the efforts of many individuals. While their activities often go unrecognized, project accomplishments can be directly attributed to their efforts. Worthy of special mention were the planning and supervisory efforts of Mary M. Hurley, Stuart G. Habley, and John A. Greene. Gail Rodabaugh and Tami Herrick were particularly helpful in the logistics efforts in past years.

The administrative and clerical chores of the project constituted a team effort of immense proportions. Assisting Carolyn Davis in the preparation of the various versions of the present document were Helen MacLean, Dorothy Reynolds, Patricia Johnson and Sara Reese. The style of this report was improved by the editorial efforts of Carol Dondrea. The production of earlier reports represented sizeable contributions by Karen Cloyd, Carolyn Davis, Carolyn Uyemura, Moira Rose, and Maryanne Leonard.

Special thanks are also due Barbara J. Rodabaugh who contributed immeasurably to many key project phases; including instrument design, literature review, program identification, and data collection. Her special contributions to the present report are reflected in the literature review section and in Appendix IV-A, "An Investigation of Longitudinal Changes in Innovative School Programs," which she authored.

Finally, our sincere appreciation is extended to the USOE Project Officer, George W. Mayeske. His continuing technical counsel, his enthusiastic support, and above all, his understanding and patience is gratefully acknowledged.

Albert B. Chalupsky
Project Director

EXECUTIVE SUMMARY

Background

During the late 1960s and early 1970s hopes were very high that intensive, innovative educational programs would demonstrate a dramatic impact on student achievement. While these hopes have moderated in recent years, there is still the feeling among many individuals, government officials as well as the general public, that if schools undergo major change they will somehow do much better. As a result, many efforts continue to be sponsored at the federal and state levels to encourage such change, either through the development of new programs or through increased emphasis on the dissemination/diffusion of innovations by educational "change agents." The wholesale adoption of educational innovations, however, is an expensive process, and there is no guarantee that the added costs will necessarily be justified by improved program effectiveness.

Educators and noneducators alike have shown a growing awareness of the lack of--and need for--evidence as to whether or not innovative educational practices are indeed better than the more traditional approaches. In response to this need, the U. S. Office of Education in 1969 awarded a contract to the American Institutes for Research to develop a design for a study of the overall effectiveness of highly intensive, innovative educational practices on students in grades 1 through 12. The initial design was developed, implemented on a limited basis, and modified during the 1969-70 school year. Full implementation began during the 1970-71 school year.

Project Objectives

The specific objectives of Project LONGSTEP (the Longitudinal Study of Educational Practices) were as follows:

- to design a system for studying the educational environments of students in terms of meaningful dimensions which underlie various educational approaches or "treatments";
- to establish a large-scale data base consisting of treatment characteristics, student characteristics, and educational outcomes

for a select sample of educational programs involving intensive and highly innovative educational practices;

- to determine, as comprehensively as possible over a three-year period of time, the impact of intensive innovation upon student performance on standardized achievement tests and on measures of educationally relevant attitudes; and
- to attempt to identify the dimensions of educational components present in a select sample of highly intensive, innovative programs that exhibit the greatest impact on student outcomes.

For purposes of this study, "intensive innovation" meant the implementation of a new program encompassing a significant proportion of students, entailing a major alteration of school procedures, and involving a high investment of resources.

Study Approach

The general emphasis of Project LONGSTEP was on the identification of changes in student achievement that occur as a result of intensive educational innovation. Highlights of the methods used to identify and document innovations and to assess student achievement are discussed next.

Selection of Schools and Students

During the early stages of the study, a nationwide search was undertaken to locate educational programs incorporating intensive, innovative practices. Guidelines for program selection emphasized program scope and intensity, instructional content, anticipated program continuation, and willingness to cooperate in a multiyear study.

Through a combination of extensive literature search, interviews with program staff, outside consultant review, and site visiting, schools located in 14 school districts were selected and agreed to participate. Of these 14 districts, 13 participated throughout the entire implementation phase and provided the data on which this report is based. Schools and districts were located in nine states: California, Florida, Kentucky, Minnesota, Pennsylvania, Texas, Utah, Washington and Wisconsin. Altogether, some 30,000 students, 80 schools, and 1,500 teachers participated in the study during its three years of implementation.

At the time of this study, most of the participating educational programs were supported largely by local funding. However, many had their roots in educational research and development activities at least partially funded by the federal government. Most of these innovations had originally been funded under Title III of the Elementary and Secondary Education Act (ESEA) of 1965. This title, "Supplementary Educational Centers and Services," was designed to encourage innovative and exemplary applications of new educational knowledge. A number were also funded by ESEA Title I, "Special Programs for Educationally Deprived Children." Combinations of funding sources were also common. One district had received support from "The Educational Professions Development Act" (EPDA) and ESEA Title II, "School Library Resources, Textbooks, and Other Instructional Materials," in addition to support from Titles I and III to fund their innovative programs. Another district was funded by ESEA Title VIII, "Dropout Prevention," in combination with Titles I and II. It should be noted, however, that the source of funding for a program was not a criterion for participation in this study and no special attempt was made to gather information on funding sources.

Sample Description

The 13 districts participating in Project LONGSTEP provided variability along various dimensions, as illustrated by a number of school and community characteristics noted during the 1970-71 school year. The communities served by the school districts ranged from 2,500 to over 600,000 in population and varied from rural to urban-metropolitan in setting. Their diversity in socioeconomic level is notable, as evidenced by the percentage of students in the participating schools receiving free or reduced-price lunches. This number ranged from less than 1 percent to nearly 30 percent. The reported instructional cost per pupil in the participating school districts varied from a low of $540 to a high of $1,050. Another indicator of the diversity in the sample was the percentage of minority group students in the participating schools; this ranged from less than 1 percent to over 30 percent.

The educational innovations encompassed by Project LONGSTEP were those that were of particular concern during the late 1960s and early 1970s--most are still of concern today. These innovations included team teaching, multimedia emphasis, unique school design, use of paraprofessionals, variations in scheduling, and teacher-developed materials, as well as independent study, student selection of materials, and a number of other practices typically associated with individualized instruction. An overview of the major educational features present in our sample at the time programs were selected, as well as the distribution of such features across districts, is presented in Table 1.

Caution should be exercised in drawing conclusions from this table, however, since the same label may have encompassed significantly different activity elements from one school district to another. By the same token, different labels were often applied to very similar activities. Further, it should be stressed that the educational activities listed in the table were often given these designations by the school systems. They did not necessarily reflect the underlying educational processes that were the primary concern of this study. Nevertheless, Table 1 does provide a gross indication of the wide variation that existed in the Project LONGSTEP sample.

Data Collection Instruments and Schedule

The data collection instruments used in this study provided information on student cognitive performance, student characteristics, educational experiences, and teacher characteristics. Student cognitive achievement was measured by either of two commercially developed instruments, the Comprehensive Tests of Basic Skills (CTBS) and the California Short Form Test of Mental Maturity (CTMM), 1963 revision. The CTBS was used for spring testing in grades 2-12 and the CTMM for testing in grade 1. Student and teacher background characteristics and attitudes were assessed by questionnaires developed specifically for Project LONGSTEP.

In order to investigate the relationships between educational practices and educational outcomes in a very diverse group of schools and programs, a system was needed for describing and quantifying the educational

TABLE 1

Overview of Major Educational Features Occurring in One or More Schools of the Districts Participating in Project LONGSTEP

Major Educational Features	School Districts												
	1	2	3	4	5	6	7	8	9	10	11	12	13
Small-group activities	X	X	X	X				X	X	X		X	X
Team teaching	X	X	X	X	X	X	X	X	X				
Multimedia emphasis	X	X	X	X	X	X	X	X	X	X		X	
Ungraded curriculum		X	X	X	X		X	X	X				
Large-group instruction		X	X					X	X	X	X	X	X
Independent study	X	X	X	X	X			X		X		X	X
Unique school design		X	X				X	X					
Differentiated staffing	X	X				X							
Volunteer aides	X	X	X	X	X	X	X	X	X	X			
IPI reading/math		X			X	X			X				
Project PLAN							X		X				
Individualized instruction (other than PLAN or IPI)	X	X		X		X	X	X	X	X		X	X
Student selection of materials	X	X	X			X		X	X	X			
Modular scheduling	X	X	X	X				X					
Teacher sensitivity training						X							
Intensive guidance and counseling						X							
Home visitation						X							
Traditional/conventional classrooms	X		X	X	X	X	X	X	X	X	X	X	X
Teacher-developed curriculum materials	X	X	X	X		X		X	X	X	X	X	X

experiences that each student had during a school year. It became evident early in the development of this system that, although there are a limited number of labels used by school districts to describe their educational activities, in practice there is a wide disparity among educational approaches sharing the same descriptive label. For example, "modular scheduling" is a commonly-used term for what are in fact various configurations of class schedules. At the same time, programs bearing different labels often turn out to be very much alike. Thus, one district's program of "individualized instruction" could be much the same as another district's "nongraded" program. Therefore, an Educational Experience Analysis Guide was developed so that complex educational experiences could be described and quantified with respect to specific observable characteristics (rather than on the basis of variously defined local labels). AIR staff used this guide to document the basic educational attributes of the school programs in which participating students were enrolled. Information on educational experiences was gathered from interviews with principals and teachers, from classroom observations, and from existing school documentation.

The guide made it possible to locate the educational experiences of participating students on a continuum ranging from traditional to innovative. Use of the guide also ensured that all students in a school who were receiving essentially the same basic educational experiences would be identified as belonging to the same "educational experience group" and be distinguished from students receiving different educational experiences, even though both groups of students may have been participating in the same school "program."

During the 1970-71 school year, this documentation methodology led to the identification of 141 educational experience groups; during the 1971-72 and 1972-73 school years, respectively, 167 and 228 groups were isolated. These groups represented all the different kinds of basic educational experiences provided by the schools for students participating in Project LONGSTEP. An illustration of the diversity of our sample is shown by the characteristics of the five school districts that contained language arts groups for the third grade during the 1972-73 school year. Figure 1 shows the average scores for these five districts on ten key educational

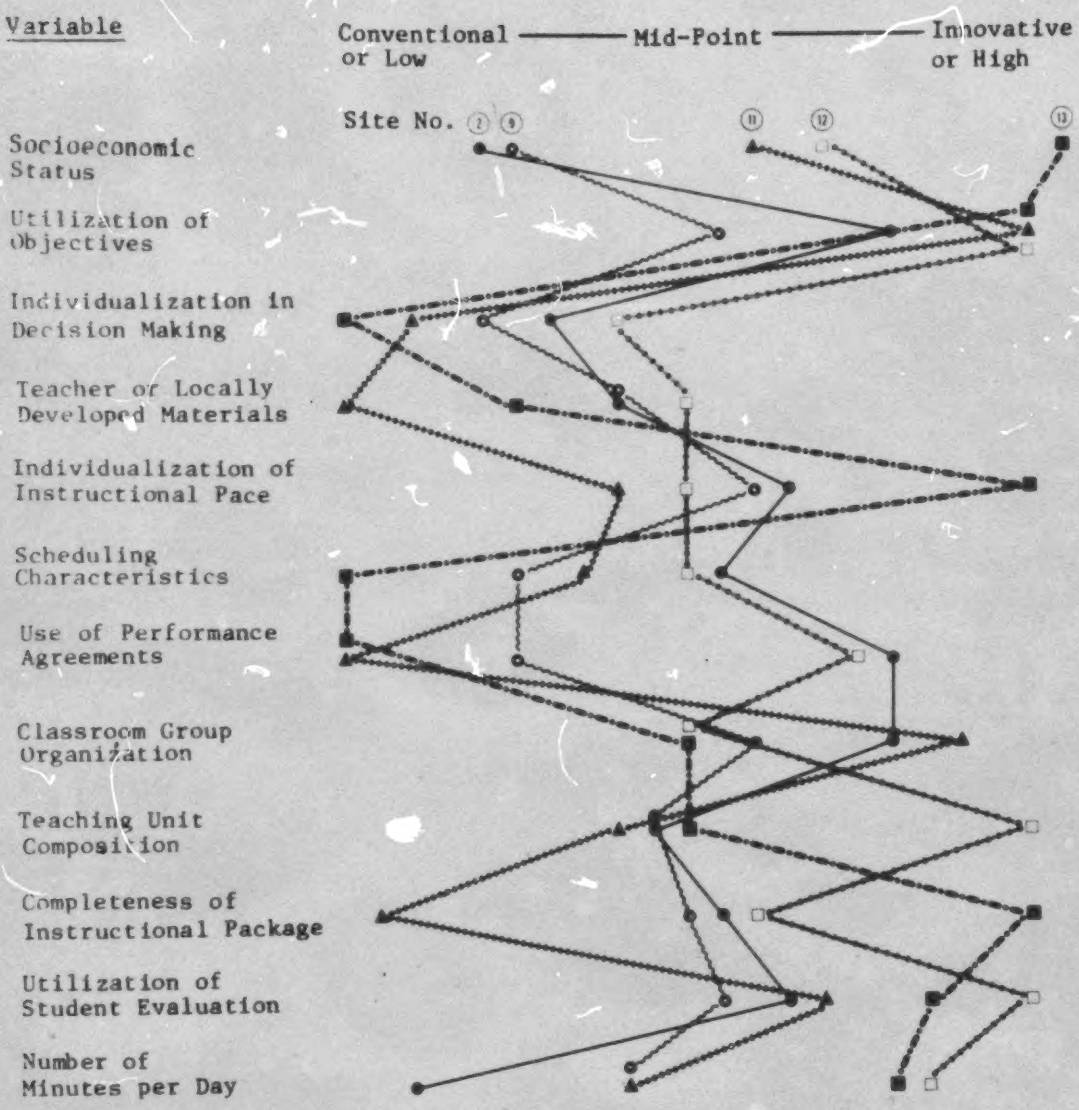


Figure 1. Profiles for language arts groups for the five districts/sites containing third grade students during the 1972-73 school year.

indices (derived from items on the Educational Experience Analysis Guide), plus student socioeconomic status, and the number of minutes per day of class time spent on language arts activities.

One final data collection guide, a classroom observation instrument, was developed to document classroom characteristics such as physical environment, study arrangements, and access to resources, as well as teacher and student activities such as degree of grouping, focus of activities, and use of materials.

In order to collect all the necessary data on students, teachers and educational experiences, AIR staff members typically visited each site three times during a school year. Site visits generally ranged from five to eight days in length. Documentation of educational experiences was usually accomplished during the initial visit each year.

Overview of the Data Base

Project LONGSTEP represents a pioneering effort in educational evaluation. It is perhaps the first large-scale, longitudinal study to associate the educational attributes underlying school programs and the characteristics of teachers with only those individual students who were directly exposed to their influence. The vast amount of data collected is also notable--other studies may have gathered information on more students, but for sheer quantity and complexity of data, this study is unique. When all the tests and questionnaires administered to students and teachers and all the Educational Experience Analysis Guide data collected on language arts, mathematics, social studies and science treatments over a three-year period are considered, some 3,500 individual items of data could be associated with a single student. When the scales or indices derived from all instruments are also considered, the total number of variables could increase to approximately 4,000 per student.

Because of the potential of such a large and detailed data base, a series of generalized data files and codebooks are being produced for the U.S. Office of Education so that these data can be made available to other researchers.

Scope of This Report

The present report did not attempt an exhaustive analysis of all the data collected during the three implementation years of the study. Rather, the analysis focused on the extent to which intensive innovation was associated with substantial differences in student outcomes.

Because of time and cost constraints, the analyses conducted for this report involved only three groups of students: those who started out as first graders in 1970-71, those who started out as fourth graders in 1970-71 and those who started out as sixth graders in 1970-71. These grade levels were selected for initial analysis primarily because their data would permit the examination of educational growth occurring during the primary years (grades 1 to 3), the elementary years (grades 4 to 6), and the transition years from elementary to middle or junior high school (grades 6 to 8). A future report will provide the findings for the remaining students.

The analytic scope of this report was also limited to the relationship between innovation in language arts or mathematics and achievement in reading, language and arithmetic. An analysis of the generalized impact of innovation throughout the entire school environment (reflecting educational experiences in social studies and science as well as language arts and mathematics) will be presented in a future report.

The measures that were related to posttest achievement included each student's socioeconomic status (SES) and pretest score; the educational attributes to which each student was exposed, including the overall amount of innovation and individualization in the school program; the amount of class time allocated by that program to language arts or mathematics instruction; and the qualifications of each student's teachers. Amount of innovation was measured by an index called Level of Innovation. This scale was equal to the sum of the following scales derived from the Educational Experience Analysis Guide: Individualization in Decision Making, Individualization of Instructional Pace, Use of Performance Agreements, Utilization of Student Evaluation, Utilization of Objectives, Teacher or Locally Developed Materials, Scheduling Characteristics, Classroom Group

Organization, Teaching Unit Composition, and Completeness of Instructional Package. Our measure of program emphasis on individualized instruction, Degree of Individualization, was defined as the sum of only the first four of these scales.

Although Project LONGSTEP collected data for three consecutive school years, two pretest/posttest analyses (Year 1 versus Year 2 and Year 2 versus Year 3) were viewed as the most powerful approach for the initial analysis of these data. Because of the multilevel nature of the study's data base, project staff concluded that overly complex analytic models should not be utilized until the data and the relationships among the basic variables were more clearly understood.

In order to determine whether intensive innovation was related to substantial gains in achievement in Project LONGSTEP's diverse sample of schools, the study assumed that if innovation does have a dramatic impact on achievement, then:

- variation with respect to intensity of innovation, even within a sample of generally innovative schools, should be positively correlated with achievement;
- such impact should be a general trend across different classrooms, schools and school districts and should be consistently present in consecutive school years;
- the achievement gains demonstrated by Project LONGSTEP's generally innovative sample of schools should noticeably exceed the gains shown by national norms; and
- such impact should be large enough to be demonstrated by a number of different methods, including those involving the kinds of statistical adjustments normally required in analyzing nonexperimental data.

Because Project LONGSTEP collected data from students in ongoing school programs, it was deemed essential that the study's key hypotheses be evaluated from a number of slightly different methodological perspectives so as to minimize the possibility that findings would be highly method- or assumption-dependent. Conclusions could then be based on converging lines of evidence. For this reason, the analyses implemented for this report were, by design, as intensive as possible within project time and cost constraints.

Findings in Review

So that they might be summarized in a concise and meaningful way, the Project LONGSTEP findings have been organized into four sections. Each section is headed by the overall, basic question to which the findings pertain.

Were Students of Different Socioeconomic Status and Pretest Levels Exposed to Different Kinds of Treatments?

In a number of grades, lower SES and lower pretest students were members of schools and instructional programs that spent more time on language arts and mathematics activities. There also was a very slight tendency across all analyses for lower SES students to be exposed to less innovative and less individualized programs of instruction. Even the most notable of these trends, however, was extremely small relative to the magnitude of the differences in socioeconomic status among the groups studied. The general finding that student SES and pretest background were not both highly and systematically related to the characteristics of educational approaches means that it is not likely that such factors biased (in a constant manner across all analyses) the overall results of Project LONGSTEP concerning the impact of intensity of innovation.

To What Extent Is Achievement Related to Pretest and Socioeconomic Status?

Both initial achievement level (i.e., pretest) and student socioeconmic status (SES) are usually substantially related to posttest achievement in school effects studies. The results of Project LONGSTEP are no exception. It is notable, however, that:

- pretest was a vastly better predictor of posttest achievement than was SES;
- the predictive value of SES was reduced almost to zero after the predictive overlap with pretest was taken into account; and,

- the relationship between pretest and posttest was still very large even after student differences with respect to SES and after group differences with respect to mean pretest level were taken into account.

While not the major purpose of the study, these findings have implications for both educators and researchers. Based on a diverse sample of schools, such findings suggest that when pretest is known, it is not necessary to use socioeconomic status as a predictor of individual student posttest achievement. They also imply that SES cannot be expected to serve as a reasonably accurate substitute measure of initial achievement (i.e., pretest) level.

Did Intensive Educational Innovation Result in Substantial Growth in Achievement?

Project LONGSTEP findings with respect to the impact of educational innovation can be organized into two general categories. The first group of findings evaluated the overall amount of educational growth demonstrated by each of the samples analyzed and compared this growth with that of the CTBS norm sample. The second group of findings related variation in intensity of innovation to variation in achievement growth.

Growth in achievement and norm comparisons. Comparisons among average pretest performance, average posttest performance, and national norms showed that:

- the greatest gains in reading, language and arithmetic achievement occurred in the earlier grades;
- the magnitude of yearly gains in achievement tended to decrease with each higher grade;
- for all grades, achievement gains seemed to be slightly larger in arithmetic than in language arts (reading and language);
- although gains were not dramatic, achievement growth during third grade was somewhat greater than that in the norm sample; and

- Project LONGSTEP students, on the average, did not do conspicuously better on posttest (relative to national norms) than they did on pretest.

In summary, the expectation that substantial yearly gains in student achievement would occur for a sample of intensive, innovative educational programs is not supported by these findings.

Growth in achievement and innovative emphasis. The most acceptable approach to analyzing data obtained by nonexperimental/associational studies (like Project LONGSTEP) is to use a number of different analysis methods and look for consistent results across methods. Therefore, Project LONGSTEP analyzed:

- overall differences in achievement growth among educational approaches;
- an educational growth model in which achievement was related to innovative emphasis, pretest, socioeconomic status (SES), and teaching qualifications;
- a posttest achievement score statistically adjusted for pretest and SES differences;
- groups of students with similar educational experiences who, on the average, performed either much better or much worse than was expected from their pretest and SES; and
- consistently overachieving and underachieving students.

The major findings based on these methods are reported next with respect to reading and arithmetic achievement. No separate discussion is presented for language achievement since these findings closely paralleled those for reading. A later section contains a more general summary of project findings.

Reading achievement. The analyses of reading achievement indicated that:

- meaningful but not dramatic differences in educational growth did exist among different educational experience groups, especially those involving students in the early elementary grades;

- differences among educational experience groups decreased with each higher grade;

- in spite of group differences in achievement, no consistent overall relationship existed between innovative intensity (as indexed by the study's measure of general innovativeness, Level of Innovation) and posttest reading performance, or between individualization emphasis (as indexed by Degree of Individualization) and posttest reading performance;

- in the only series of analyses in which growth in reading achievement was related to overall innovation and individualization to a substantial degree (during the third grade), the impact of Level of Innovation (or Degree of Individualization) was negative--that is, the greatest educational growth occurred in programs with a more moderate emphasis on innovation;

- the study's measure of teacher experience, Teaching Qualifications, was not related to posttest performance to any meaningful degree in our analyses; and

- the analysis of unusually effective reading groups provided no clear evidence that Level of Innovation or Degree of Individualization was consistently related to achievement. This was also true of the analysis of students who demonstrated significant overachievement or underachievement during two consecutive school years. In fact, the consistent overachievers actually tended to be members of programs with a lower Level of Innovation or Degree of Individualization.

An additional finding regarding the study's measure of amount of schooling, number of minutes per day of class time on language arts, is noteworthy. Students who exhibited unusually large gains in reading achievement during two consecutive school years were exposed to much more language arts instruction as second graders than were students with a notable lack of growth (i.e., 113 minutes per day compared with 85 minutes). The consistent overachievers, as third graders, were exposed to about 20 minutes of instruction per day less than during the previous school year, and yet they again demonstrated larger than expected gains in achievement. These results suggest that increased exposure to language arts during the second grade may significantly improve the chances of some students to demonstrate substantial gains in reading achievement, even in later grades.

Arithmetic/Mathematics achievement. Meaningful differences among educational experience groups existed with respect to arithmetic posttest achievement, especially at the lower grade levels. In fact, differences among these groups decreased with each higher grade. Other analyses showed that:

- Level of Innovation or Degree of Individualization was negatively, rather than positively, related to growth in arithmetic achievement. This appeared to be the most notable result in the samples as a whole. Therefore, the hypothesis that greater overall emphasis on innovation or individualization is positively and consistently related to student achievement in arithmetic was clearly not supported.
- Neither Level of Innovation nor Degree of Individualization demonstrated a dramatic or consistently positive effect on student arithmetic achievement in (1) the analyses of unusually effective educational experience groups and (2) the analyses of students who demonstrated significant overachievement or underachievement for two consecutive school years. In fact, relative to underachieving students, the students who were identified as overachievers, on the average, were members of programs with a more moderate emphasis on innovation or individualization.

Overview of achievement and innovative emphasis. In summary, no evidence could be found that either of the major treatment variables--Level of Innovation and Degree of Individualization--was substantially and positively correlated with posttest performance. Further, preliminary analyses showed that, on the average, these findings applied equally well to students at different SES or pretest levels.

Meaningful differences in the teacher qualifications of groups performing at a higher or lower level than predicted were infrequent. Whenever they did occur, however, better teaching qualifications (as indicated by such items as education, teaching experience, certification and type of appointment) were associated with better student performance. And finally, younger students (i.e., those in grades 2 and 3) who achieved substantially more than expected during two consecutive years were exposed to a notably greater amount of class time per day on language arts than were the consistent underachievers during the second grade. Even though they spent considerably less class time per day on language arts in grade 3, the overachievers again demonstrated dramatic gains.

Can Schools Have a Substantial Impact on Educational Achievement?

The findings show that substantial educational growth was not uniquely associated with innovative emphasis. Even small amounts of growth were not positively and consistently related to our measures of innovation, Level of Innovation and Degree of Individualization. It should not be concluded, however, that substantial growth was absent in the schools and students participating in Project LONGSTEP. On the contrary, as discussed previously, a number of educational experience groups were identified whose students, on the average, did much better than was predicted from their pretest and SES level. For example, as indicated in Figure 2, the average achievement gains of sixth grade students in the most effective reading treatment groups differed dramatically from those exhibited by students in the groups that did much worse than was predicted from their pretest score and their socioeconomic status. The students in the poorer performing reading groups shown in Figure 2 had average pretest scores ranging between the 56th and 74th percentiles; their average posttest scores dropped slightly, ranging between the 41st and 66th percentiles. In contrast, the sixth grade

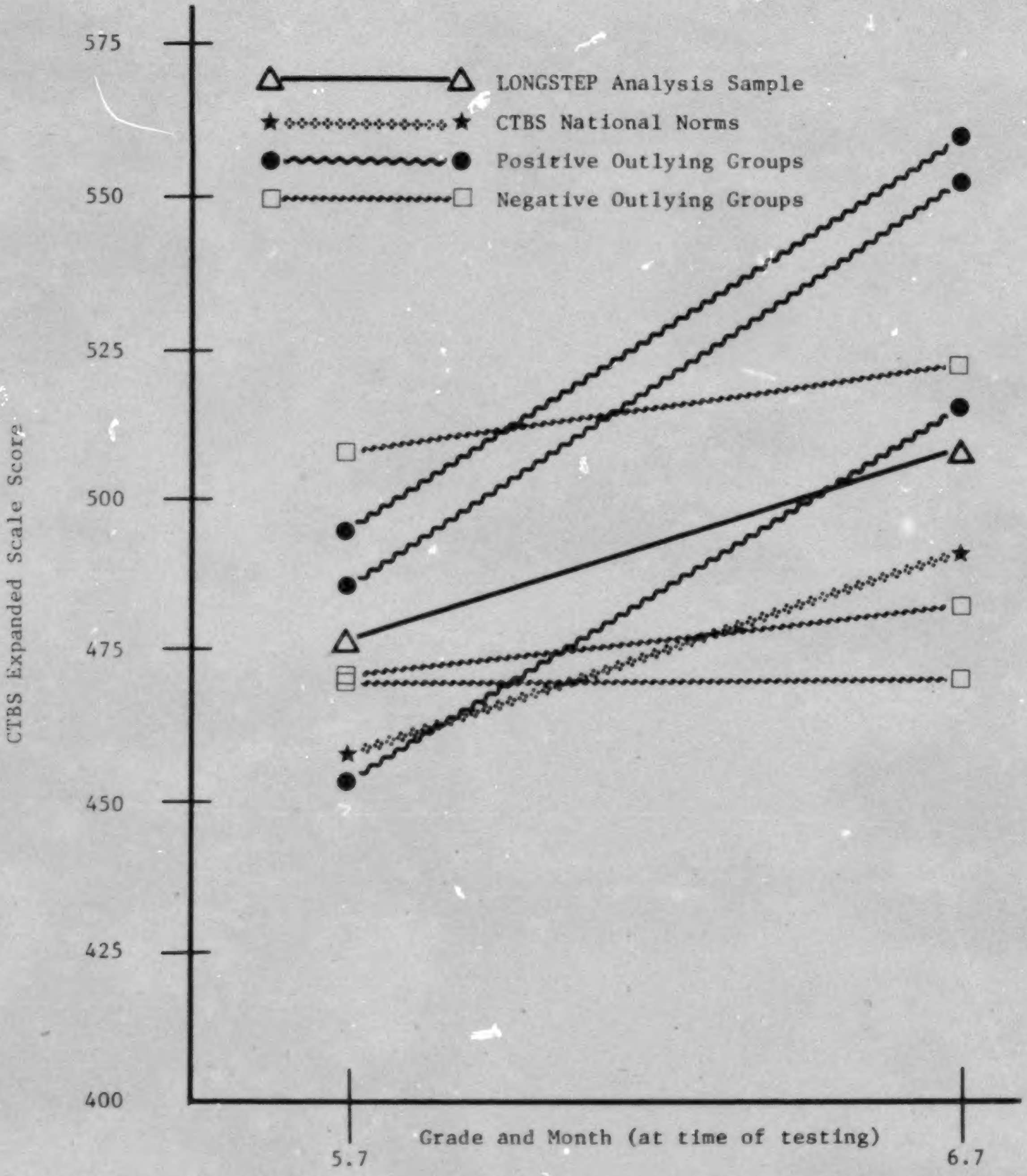


Figure 2. Educational experience groups where students achieved more or less than was expected from their pretest and SES level (Spring 1973 Reading Analysis).

students in the high performing reading groups, who had average pretest scores ranging between the 48th and 69th percentiles, improved noticeably, with average posttest scores ranging between the 63rd and 79th percentiles.

Other examples of dramatic gains in achievement were shown in the analyses identifying individual students who, for two consecutive years, demonstrated posttest achievement differing greatly from that expected on the basis of their pretest and SES scores.

The analyses of unusually effective groups and the analyses of students with unusual growth patterns both provide evidence that large gains in cognitive achievement were occurring in Project LONGSTEP. Thus, even though the overall project findings showed that dramatic school effects were not associated with intensity of innovation, the findings noted here suggest that substantial achievement gains, over and above those expected on the basis of pretest and SES, do occur in public schools. It should also be remembered that the overall analyses showed that small but meaningful posttest differences existed among our educational experience groups, although they were not uniquely associated with the project's measures of innovative intensity.

These findings certainly are not surprising to teachers, principals, parents or even researchers who have personally witnessed impressive growth in cognitive skills. They are, however, somewhat more "newsworthy" in terms of educational evaluation results. Taken together, then, these findings must lead us to conclude that:

- different educational approaches do produce meaningful and important differences in achievement (especially in the early elementary grades); and
- although substantial gains in achievement were not demonstrated by a majority of Project LONGSTEP programs/schools and students, unquestionably dramatic improvements in reading, language and arithmetic skills were found in some schools.

Future research and evaluation efforts should be directed at identifying the determinants of such significant educational gains.

Overall Conclusions and Possible Implications

The single most important and well-documented finding of Project LONGSTEP was the lack of either substantial or consistent association between student achievement and overall Level of Innovation across grades. In fact, relative to the sample of schools participating in the project, students enrolled in programs with a more moderate emphasis on innovation and individualization showed the greatest improvement. This finding should perhaps not be viewed as surprising or regrettable. Why should change, or innovation, automatically mean improvement? On the contrary, it is realistic to expect--and the Project LONGSTEP data supported this expectation--that some changes will be improvements, some changes will be detrimental, and some changes will not matter at all.

From the standpoint of time spent per day on language arts activities, it appeared that consistently overachieving students in Project LONGSTEP typically were members of programs that allocated more time to such instruction in the second grade. These findings suggest that increasing the amount of class time per day, especially in the first and second grade, may be a worthwhile strategy for improving student performance in language arts.

Our various analyses also suggest that the impact of the educational practices studied was inversely related to grade level. Specifically, it appeared that differences among educational experience groups, independent of the effects of student SES level, initial achievement level, and Teacher Age, SES Background and Teaching Qualifications, were greatest for second and third graders, less for fifth and sixth graders, and almost nonexistent for seventh and eighth graders. These findings tend to support the view that the greatest potential payoff from the investment of future educational research and instructional development funds may come from increased attention to improving education in the early elementary grades. It is suggested that further effort be expended to gain a better understanding of the educational practices that show the greatest benefit at these early grades. The results of such research could then serve as the basis for designing educational strategies that could have a relatively greater overall impact on American education.

It is less obvious why one particular aspect of much recent innovation, that is, increased individualization of education, has not been shown to be clearly and consistently effective in Project LONGSTEP. Some possible explanations for this finding will be discussed, but it should be emphasized that these are purely speculative. One possibility is that not all modes of "individualization" operate effectively and that our index, Degree of Individualization, failed to tap some crucial differences between individualized programs that are effective and those that are not. The program characteristics measured by this index included four scales: Individualization in Decision Making (covering decisions on topic selection, materials selection, and sequencing decisions); Individualization of Instructional Pace; Use of Performance Agreements; and Utilization of Student Evaluation (covering the extent to which tests are used for evaluating student needs, and whether or not such evaluation serves as a basis for modifying a student's instructional program).

The extent to which the level of performance is left to the student could be a crucial factor. For example, it may be that tailoring the educational process or teaching strategy to an individual student's needs or learning style can be very effective but that allowing the student a great deal of latitude in establishing the level or standard of attainment during a school year can have a negative impact on performance (as measured by standardized achievement tests) for all but the highly motivated. The extent to which teachers monitored, encouraged, prodded, etc., their students may well be crucial but this was not tapped by our data collection instruments. Anecdotal experiences gained over the years of association with schools participating in Project LONGSTEP suggest that some students tend to lose momentum unless the teacher provides the kind of individual attention that is unlikely except when a sufficient number of qualified assistants are available.

The fact that our Degree of Individualization index failed to discriminate between successful and unsuccessful educational approaches also suggests the possibility that we may have overlooked some elements of individualization that occur in conventional classrooms. It could be that individualization is one of the practices that distinguishes good teachers from poor ones, even in traditional educational environments where there is no "experimental" or "innovative" program claiming to be individualized. Perhaps the good teacher individualizes informally and so instinctively that he or she is not even aware of it in these terms. The good teacher does not treat all children as interchangeable cogs; but rather as the individuals they are. This kind of informal individualization would not be noted as a feature of an educational program in Project LONGSTEP--but it may nevertheless be very effective.

On the other hand, it is also possible that large overall, educational effects were not demonstrated in Project LONGSTEP (or have not been found in many educational evaluation efforts) because such effects have been attenuated by inappropriate matches between educational approaches and student needs. The undoubtedly dramatic growth in achievement demonstrated by a number of students participating in Project LONGSTEP certainly suggests that some near-optimal match of student and educational approach may have been one of the reasons for the gains of these students.

In final summary, the Project LONGSTEP findings should not be viewed as a sweeping criticism of the concepts of innovation or individualization. However, they should serve as a reminder to educators--as well as to parents and legislators--that educational innovation *per se* will not necessarily produce dramatic effects on student achievement. Educational quality is *not* synonymous with innovation or individualization. Individualization of instruction may represent a valuable approach for the improvement of American education, but the findings here suggest that individualization as a program strategy should not be viewed as the final or complete answer. In this case, more (innovation) is not necessarily better (education).

TABLE OF CONTENTS

PREFACE

EXECUTIVE SUMMARY

Page

I. INTRODUCTION 1

A. Historical Perspective 1

B. Study Objectives 3

C. Related Research 4

School Impact 4

Studies of School Processes 11

II. TECHNICAL APPROACH 15

A. Introduction 15

B. Selection of Programs, Schools and Students 16

Program Selection 16

Description of Participating Programs 21

Selection of Schools and Students 24

C. Data Collection Instruments and Procedures 27

Student Background and Outcome Measures 28

Teacher Questionnaires 33

Pretest and Revision of the Student Questionnaires and the Teacher Questionnaire 35

Educational Treatment Documentation 36

Classroom Documentation 39

Class Membership Forms and Class Rosters 41

Student Questionnaire Verification 42

Data Collection Schedule 43

D. Data Management and Quality Control Procedures 45

Overview of the Magnitude of the Project LONGSTEP Data Base 45

Project LONGSTEP's Data Management Approach 46

III. DISCUSSION OF VARIABLES EXAMINED IN THIS REPORT 57

A. Introduction 57

B. Standardized Achievement Test Variables 59

TABLE OF CONTENTS (continued)

Page

C. Questionnaire Scale Analysis Procedures and Results . . . 67
 Introduction 67
 Procedures and Results 67

D. Student Questionnaire Verification Study 79

E. Educational Experience Scale Analysis Procedures and Results 80

F. Measures of Treatment Exposure 88

G. Classroom Observation Scale Analysis Procedures and Results 90

H. Analytic Scope of This Report 92

I. Overview of the Analysis Data Base 93

IV. THE ANALYSIS STRATEGY 95

A. Introduction 95

B. Analytic Questions and Strategies 96

V. RESULTS AND DISCUSSION 143

A. Introduction 143

B. The Analysis of Reading Achievement 145
 Descriptive Overview of the Analysis Samples 145
 Relationships among Treatments, Pretest and SES 152
 Relationships of Key Analysis Variables with Posttest Reading Performance 163
 Analysis of the Two Revised Growth Models 180
 Examination of Unusually Effective EdExAG Groups . . . 196
 Examination of Unusual Student Growth Patterns in Reading Achievement 207

C. The Analysis of Language Achievement 222
 An Overview of Language Achievement 222
 Examination of Unusually Effective EdExAG Groups . . . 224
 Examination of Unusual Student Growth Patterns in Language Achievement 225

TABLE OF CONTENTS (continued)

Page

D. The Analysis of Arithmetic Achievement 229
- Descriptive Overview of the Analysis Samples 229
- Relationships among Treatments, Pretest and SES . . . 235
- Relationships of Key Analysis Variables with Posttest Arithmetic Performance 244
- Analysis of the Two Revised Growth Models 260
- Examination of Unusually Effective EdExAG Groups . . . 273
- Examination of Unusual Student Growth Patterns in Mathematics Achievement 282

VI. SUMMARY AND CONCLUSIONS 297

A. Study Objectives 297

B. Sample Description 298

C. Documentation of Educational Practices 299

D. Scope of This Report 302

E. Findings in Review 303
- Were Students of Different Socioeconomic Status and Pretest Levels Exposed to Different Kinds of Treatments? 306
- To What Extent Is Achievement Related to Pretest and Socioeconomic Status? 307
- Did Intensive Educational Innovation Result in Substantial Growth in Achievement? 308
- Can Schools Have a Substantial Impact on Educational Achievement? 317

F. Study Limitations 320

G. Overall Conclusions and Possible Implications 327

REFERENCES 333

APPENDICES (Bound in a Separate Volume)

LIST OF TABLES

Table No. Pages

II-1 Participating School/Community Characteristics Based on Data Obtained during 1970-71 School Year 22-23

II-2 Overview of Major Educational Features Occurring in One or More Schools of the Districts Participating in Project LONGSTEP 25

III-1 Student Achievement Test Variables 1970-71 60

III-2 Student Achievement Test Variables 1971-72 60

III-3 Student Achievement Test Variables 1972-73 61

III-4 Published Reliability Coefficients (Kuder-Richardson Formula #20) for Selected Subtest Scores 62

III-5 Adjusted Internal Consistency Reliability Coefficients for the CTBS Form Q 64

III-6 Proportion of School Year Prior to Test Administration (By Site) 66

III-7 Student Questionnaire Scales, Number of Items, Scale Item Composition and Item Numbers 71-72

III-8 Teacher Questionnaire Scales, Number of Items, Scale Item Composition and Item Numbers 73-74

III-9 Coefficient Alphas for All Multi-item Student Questionnaire Scales 78

III-10 Coefficient Alphas for All Multi-item Teacher Questionnaire Scales 79

III-11 Educational Experience Analysis Guide Scales, Number of Items, Scale Item Composition and Item Numbers . . 83-86

III-12 Coefficient Alphas for All Multi-item Educational Experience Analysis Guide Scales 87

III-13 Measures of Treatment Exposure and Staff Resources 89

III-14 Classroom Observation Scales, Number of Items, Scale Item Composition and Item Numbers 91

III-15 Coefficient Alphas for All Multi-item Classroom Observation Form Scales 92

III-16 Grade-level Membership for Students in Cohorts 1, 4 and 6 93

LIST OF TABLES (continued)

Table No.		Pages
IV-1	Numbers of Students in Each Analysis Sample for Cohorts 1, 4 and 6	103
IV-2	Means (μ') and Variance (σ'^2) of a Truncated Normal Distribution ($\mu = 0$, $\sigma^2 = 1.0$)	107
IV-3	Means, Standard Deviations and Point-Biserial Correlations (r_{pb}) for the Missing Data Analysis of Cohort 1, Spring 1973 Reading Total Score Analysis Sample	109
IV-4	Means, Standard Deviations and Point-Biserial Correlations (r_{pb}) for the Missing Data Analysis of Cohort 4, Spring 1973 Reading Total Score Analysis Sample	110-111
IV-5	Means, Standard Deviations and Point-Biserial Correlations (r_{pb}) for the Missing Data Analysis of Cohort 6, Spring 1973 Reading Total Score Analysis Sample	112-113
IV-6	Sex and Ethnic Group Membership for the Cohort 1, 4 and 6 Analysis Samples and Missing Data Samples	116
IV-7	Key Predictor Variables	120
IV-8	Posttest Means, Standard Deviations, Numbers of EdExAG Groups and Students at Each Level of the Individualization of Instructional Pace (IIP) Scale (Cohort 1, Spring 1972 Reading Total Score Analysis)	125
IV-9	Multiple Regression Analysis of the Spring 1972 Reading Total Criterion--Cohort 4 (N = 1952; n = 42)	131
V-1	Descriptive Summary of the Reading Total Score Analysis Samples	146-147
V-2	Mean Differences between Reading Total Pretest and Posttest Scores and "Standardized" Gains	148
V-3	Proportions of SES and Pretest Variance Attributable to EdExAG Group Membership	153
V-4	Correlations of SES with the Key Analysis Variables - Reading Total Score Analysis	157-158
V-5	Correlations of Pretest with the Key Analysis Variables - Reading Total Score Analysis	159-160

LIST OF TABLES (continued)

Table No. Pages

V-6 Commonality Coefficients for the Regression of Posttest on SES (S), Pretest (P), EdExAG Dummy Codes (D) and Teacher Variables (T)--Cohort 1, Spring 1972, Reading Total Score Analysis 165

V-7 Commonality Coefficients for the Regression of Posttest on SES (S), Pretest (P), EdExAG Dummy Codes (D) and Teacher Variables (T)--Cohort 1, Spring 1973, Reading Total Score Analysis 166

V-8 Commonality Coefficients for the Regression of Posttest on SES (S), Pretest (P), EdExAG Dummy Codes (D) and Teacher Variables (T)--Cohort 4, Spring 1972, Reading Total Score Analysis 167

V-9 Commonality Coefficients for the Regression of Posttest on SES (S), Pretest (P), EdExAG Dummy Codes (D) and Teacher Variables (T)--Cohort 4, Spring 1973, Reading Total Score Analysis 168

V-10 Commonality Coefficients for the Regression of Posttest on SES (S), Pretest (P), EdExAG Dummy Codes (D) and Teacher Variables (T)--Cohort 6, Spring 1972, Reading Total Score Analysis 169

V-11 Commonality Coefficients for the Regression of Posttest on SES (S), Pretest (P), EdExAG Dummy Codes (D) and Teacher Variables (T)--Cohort 6, Spring 1973, Reading Total Score Analysis 170

V-12 Correlations of All Key Analysis Variables with Posttest - Reading Total Score Analysis 176-177

V-13 Correlations of the Key Treatment Variables with Posttest Residualized on the Basis of Pretest and SES - Reading Total Score Analyses 179

V-14 Commonality Coefficients for the Level of Innovation Growth Model--Cohort 1, Spring 1972, Reading Total Score Analysis 182

V-15 Commonality Coefficients for the Level of Innovation Growth Model--Cohort 1, Spring 1973, Reading Total Score Analysis 183

V-16 Commonality Coefficients for the Level of Innovation Growth Model--Cohort 4, Spring 1972, Reading Total Score Analysis 184

V-17 Commonality Coefficients for the Level of Innovation Growth Model--Cohort 4, Spring 1973, Reading Total Score Analysis 185

LIST OF TABLES (continued)

Table No. Pages

V-18 Commonality Coefficients for the Level of Innovation Growth Model--Cohort 6, Spring 1972, Reading Total Score Analysis 186

V-19 Commonality Coefficients for the Level of Innovation Growth Model--Cohort 6, Spring 1973, Reading Total Score Analysis 187

V-20 Intercorrelations among the Level of Innovation Model Variables for the Cohort 1, Spring 1973 Analysis Sample . 189

V-21 Regression Coefficients for the Level of Innovation Growth Model - Reading Total Score Analyses 191-192

V-22 Regression Coefficients for the Degree of Individualization Growth Model - Reading Total Score Analyses . . . 193-194

V-23 Comparison of Key Analysis Variables for Positive and Negative Outlier EdExAG Groups--Cohort 1, Spring 1972, Reading Total Score Analysis 198-199

V-24 Comparison of Key Analysis Variables for Positive and Negative Outlier EdExAG Groups--Cohort 1, Spring 1973, Reading Total Score Analysis 200-202

V-25 Comparison of Key Analysis Variables for Students Exhibiting Unusual Longitudinal Growth Patterns In Two Consecutive School Years--Cohort 1, Reading Total Score Analysis 208-210

V-26 Comparison of Key Analysis Variables for Students Exhibiting Unusual Longitudinal Growth Patterns In Two Consecutive School Years--Cohort 4, Reading Total Score Analysis 213-215

V-27 Comparison of Key Analysis Variables for Students Exhibiting Unusual Longitudinal Growth Patterns In Two Consecutive School Years--Cohorts 6, Reading Total Score Analysis 217-219

V-28 Descriptive Summary of the Arithmetic Total Score Analysis Samples 230-231

V-29 Mean Differences between Arithmetic Total Pretest and Posttest Scores and "Standardized" Gain 232

V-30 Proportions of SES and Pretest Variance Attributable to EdExAG Group Membership 236

V-31 Correlations of SES with the Key Analysis Variables - Arithmetic Total Score Analysis 238-239

V-32 Correlations of Pretest with the Key Analysis Variables - Arithmetic Total Score Analysis 240-241

LIST OF TABLES (continued)

Table No. Pages

V-33 Commonality Coefficients for the Regression of Posttest on SES (S), Pretest (P), EdExAG Dummy Codes (D) and Teacher Variables (T) -- Cohort 1, Spring 1972, Arithmetic Total Score Analysis 245

V-34 Commonality Coefficients for the Regression of Posttest on SES (S), Pretest (P), EdExAG Dummy Codes (D) and Teacher Variables (T) -- Cohort 1, Spring 1973, Arithmetic Total Score Analysis 246

V-35 Commonality Coefficients for the Regression of Posttest on SES (S), Pretest (P), EdExAG Dummy Coes (D) and Teacher Variables (T) -- Cohort 4, Spring 1972, Arithmetic Total Score Analysis 247

V-36 Commonality Coefficients for the Regression of Posttest on SES (S), Pretest (P), EdExAG Dummy Codes (D) and Teacher Variables (T) -- Cohort 4, Spring 1973, Arithmetic Total Score Analysis 248

V-37 Commonality Coefficients for the Regression of Posttest on SES (S), Pretest (P), EdExAG Dummy Codes (D) and Teacher Variables (T) -- Cohort 6, Spring 1972, Arithmetic Total Score Analysis 249

V-38 Commonality Coefficients for the Regression of Posttest on SES (S), Pretest (P), EdExAG Dummy Codes (D) and Teacher Variables (T) -- Cohort 6, Spring 1973, Arithmetic Total Score Analysis 250

V-39 Correlations of All Key Analysis Variables with Posttest - Arithmetic Total Score Analysis 256-257

V-40 Correlations of the Key Treatment Variables with Posttest Residualized on the Basis of Pretest and SES - Arithmetic Total Score Analyses 259

V-41 Commonality Coefficients for the Level of Innovation Growth Model -- Cohort 1, Spring 1972, Arithmetic Total Score Analysis 261

V-42 Commonality Coefficients for the Level of Innovation Growth Model -- Cohort 1, Spring 1973, Arithmetic Total Score Analysis 262

V-43 Commonality Coefficients for the Level of Innovation Growth Model -- Cohort 4, Spring 1972, Arithmetic Total Score Analysis 263

V-44 Commonality Coefficients for the Level of Innovation Growth Model -- Cohort 4, Spring 1973, Arithmetic Total Score Analysis 264

LIST OF TABLES (continued)

Table No. Pages

V-45 Commonality Coefficients for the Level of Innovation Growth Model -- Cohort 6, Spring 1972, Arithmetic Total Score Analysis 265

V-46 Commonality Coefficients for the Level of Innovation Growth Model -- Cohort 6, Spring 1973, Arithmetic Total Score Analysis 266

V-47 Regression Coefficients for the Level of Innovation Growth Model - Arithmetic Total Score Analyses 269-270

V-48 Regression Coefficients for the Degree of Individualization Growth Model - Arithmetic Total Score Analyses . . 271-272

V-49 Comparison of Key Analysis Variables for Positive and Negative Outlier EdExAG Groups--Cohort 1, Spring 1972, Arithmetic Total Score Analysis 275-276

V-50 Comparison of Key Analysis Variables for Positive and Negative Outlier EdExAG Groups--Cohort 1, Spring 1973, Arithmetic Total Score Analysis 277-279

V-51 Comparison of Key Analysis Variables for Students Exhibiting Unusual Longitudinal Growth Patterns In Two Consecutive School Years--Cohort 1, Arithmetic Total Score Analysis 284-286

V-52 Comparison of Key Analysis Variables for Students Exhibiting Unusual Longitudinal Growth Patterns In Two Consecutive School Years--Cohort 4, Arithmetic Total Score Analysis 288-290

V-53 Comparison of Key Analysis Variables for Students Exhibiting Unusual Longitudinal Growth Patterns In Two Consecutive School Years--Cohort 6, Arithmetic Total Score Analysis 292-294

LIST OF FIGURES

Figure No.		Pages
II-1	Data collection schedule for Project LONGSTEP	44
II-2	General data management system of Project LONGSTEP . . .	54
IV-1	Variance sources for the regression of a student-level dependent variable on a quantitative EdExAC-level predictor (X)	126
IV-2	Possible growth patterns of students who obtained posttest scores which deviated from those predicted from pretest and SES	140
V-1	Pretest and posttest means and published national norms (50th percentile for spring testing) - Reading Total Score analyses	149
V-2	Pretest and posttest means and published national norms (50th percentile for spring testing) - Arithmetic Total Score analyses	233
VI-1	Profiles for language arts treatments for the five sites containing third grade students during the 1972-73 school year	301
VI-2	Cohort 4 treatment groups where students achieved more or less than was expected from their pretest and SES level (Spring 1973 Reading Analysis)	318

I. INTRODUCTION

A. Historical Perspective

During the past decade, Americans have become increasingly concerned about the effectiveness of their schools. Educators have responded to this concern largely by accelerating their search for new techniques, new materials, and new curricula that would significantly improve the achievement of students. The wholesale adoption of educational innovations, however, is an expensive process, and there is no guarantee that the added costs will be justified by improved program effectiveness. Since the citizen expects his tax dollar to buy not merely new educational programs but more effective ones, it has become increasingly more important for educators to demonstrate the efficacy of such new educational practices. The number of educational bond issues that fail to pass in each election, however, suggests that the rising cost of education has not been adequately justified.

Educators and noneducators alike have shown a growing awareness of the lack of evaluative data about which of the new methods and curricula were particularly effective. Disquieting questions have been raised over the years about whether the new educational practices were indeed "better" than more traditional approaches. Congress responded to this concern in the late 1960s by making funds available to the U. S. Office of Education (USOE) for the large-scale evaluation of educational programs. This funding resulted in such projects as the Evaluation of the Emergency School Assistance Program (RMC, Incorporated, 1971) and the Study of Selected Exemplary Programs for the Education of Disadvantaged Children (Hawkridge, Chalupsky and Roberts, 1968). The study reported here, while differing significantly from these specific program evaluations, is a further outgrowth of the commitment to evaluative research.

The Longitudinal Study of Educational Practices (Project LONGSTEP) sought to assess the overall impact of intensive, innovative educational practices or "treatments" on student outcomes, taking into account differences in student characteristics and in school environments. Two major issues stimulated this study:

1. Do intensive levels of educational innovation make a significant difference in student outcome?
2. What components or elements within intensive educational efforts significantly affect student outcome?

Previous evaluation research involving the study of classroom variables was often limited in scope and restricted in period of time covered. At the same time, large-scale evaluation studies typically dealt only with school resources and did not attempt to study the characteristics of specific educational practices. This study attempts to extend the knowledge gained from both large-scale, cross-sectional studies and from small-scale, in-depth studies of classroom variables. It attempts to overcome the weaknesses of these studies by applying the same measures of input and output to rather large samples of students, and by studying specific student characteristics and specific educational innovations over a period of several years.

The study originated in 1969 when the American Institutes for Research (AIR) was awarded a contract to develop the design for a detailed study of the effects of selected education programs on children in Grades 1 through 12. The initial design was developed, implemented on a limited basis, and modified during the 1969-70 school year. Full implementation began during the 1970-71 school year. The study represents the cooperative effort of 13 school districts throughout the United States, the Office of Planning, Budgeting and Evaluation of USOE, and AIR.

At the time of this study, most of the participating educational programs were supported largely by local funding. However, many had their roots in educational research and development activities at least partially funded by the federal government. Most of these innovations had originally been funded under Title III of the Elementary and Secondary Education Act (ESEA) of 1965. This title, "Supplementary Educational Centers and Services," was designed to encourage innovative and exemplary applications of new educational knowledge. A number were also funded by ESEA Title I, "Special Programs for Educationally Deprived

Children." Combinations of funding sources were also common. One district had received support from "The Education Professions Development Act" (EPDA) and ESEA Title II, "School Library Resources, Textbooks, and Other Instructional Materials," in addition to support from Titles I and III to fund their innovative programs. Another district was funded by ESEA Title VIII, "Dropout Prevention," in combination with Titles I and II. It should be noted, however, that the source of funding for a program was not a criterion for participation in the study and no special attempt was made to gather information on funding sources. Rather, selection of programs for the study stressed innovative practices and program intensity.

For purposes of this study, an "intensive level of effort" meant the implementation of a new program encompassing a significant proportion of students, entailing a major alteration of school procedures, and involving a high investment of resources. The overall emphasis was on the identification of changes that occur in student outcomes as a result of intensive treatment as compared with outcomes expected in the absence of intensive treatment.

By selecting for study those school systems that were implementing or about to implement large-scale innovative programs, it was hoped that educational efforts that substantially improved student performance would be discovered and documented. When a sufficient number and variety of such programs could be identified and studied in depth, it should be possible to formulate hypotheses concerning the efficacy of specific program components. Following confirmation of these hypotheses, it would then be possible for educators to design new programs having a greater probability of success than current educational systems.

B. Study Objectives

The specific objectives of the study were as follows:

1. to design a system for studying the educational environments of students in terms of meaningful dimensions which underlie various educational approaches or treatments;

2. to establish a large-scale data base consisting of treatment characteristics, student characteristics, and educational outcomes for a select sample of educational programs involving intensive and highly innovative educational practices;

3. to determine, as comprehensively as possible over a three-year period of time, the impact of intensive innovation upon student performance on standardized achievement tests and on measures of educationally relevant attitudes; and

4. to attempt to identify the dimensions of educational components present in a select sample of highly intensive, innovative programs that exhibit the greatest impact on student outcomes.

A study of the size and scope reported here involves a very large data base which can provide information on many different issues related to elementary and secondary education. The current report does not attempt an exhaustive analysis of all the data collected during the course of the study. Rather, the analysis focuses on the extent to which intensive innovation, as defined by a select set of educational practices, was associated with substantial differences in student outcomes. A discussion of these treatment variables will be presented in later chapters.

C. Related Research

Although the present study is unique in several ways, it does have antecedents in previous research. This research, discussed briefly here, is most easily categorized under two headings--studies of school impact and studies of school processes.

School Impact

Prior to the Equality of Educational Opportunity Survey (Coleman, Campbell, Hobson, McPartland, Mood, Weinfeld and York, 1966), it was generally accepted that student achievement was primarily the product of the student's school experiences. Most of the research in this category was therefore addressed to the problem of identifying significant school effects. A summary of the more comprehensive of these studies can be found in Guthrie (1970); more critical reviews have been written by Rosenshine (1970) and Spady (1973). Included in these reviews are studies that range in size from samples of 5,000 to 109,000 and cover a broad spectrum of community settings from rural to metropolitan. In spite of their diversity, these studies bear some striking similarities. In almost every case, the measure of school output is pupil performance on achievement tests, while measures of effective school service components (school input) are limited to relatively static variables. The most important of these variables, as related to student test scores, appeared to be: pupil/teacher ratio (Mollenkopf and Melville, 1956; Burkhead, Fox and Holland, 1967); instructional expenditures per student (Mollenkopf and Melville, 1956; Flanagan, Dailey, Shaycoft, Orr and Goldberg, 1962; Ribich, 1968; Benson, Schmelzle, Gustafson and Lange, 1965; and Kiesling, 1967); faculty salaries (Burkhead et al., 1967; Flanagan et al., 1962; and Benson et al., 1965); number of books in the school library (Flanagan et al., 1962; Thomas, 1962); and teacher experience (Burkhead et al., 1967; Flanagan et al., 1962). In addition, Mollenkopf and Melville reported a relationship between pupil achievement and class size, although the Flanagan study found that class size was unlikely to be a prime cause of school excellence.

The general assumption on which these studies were based--that the school has a significant impact--was challenged in 1964, when the U. S. Office of Education commissioned what was to be the most extensive and controversial educational study ever conducted in this country, the Equality of Educational Opportunity Survey (EEOS) (Coleman et al., 1966). During this survey, data were collected from more than 660,000 students in 5,000 of the nation's schools. These data included not only achievement test scores, but information about student attitudes, teacher and school characteristics, the range and diversity of school curricula, and qualifications of school administrators. Results from this study, especially the major conclusion that "schools bring little influence to bear on a child's achievement that is independent of his background and general social context," have been widely discussed and criticized during the intervening years. Although many researchers feel that the statistical methods used by Coleman and his associates placed undue significance on nonschool variables, such criticisms have tended to obscure the fact that the survey did not rule out a correlation between school variables and student achievement. However, it did place strong qualifications on such correlations. Coleman found, for example, that:

- Of the effect of variations in school facilities, curriculum and staff upon achievement, only a small part is independent of family background.
- Of the school factors, those that have the greatest influence on achievement, independent of family background, are the teachers' characteristics, not the facilities and curriculum.
- The social composition of the student body is more highly associated with achievement, independent of the student's own social background, than is any school factor.

The data bank of the Equality of Educational Opportunity Survey has been a significant resource for educational research. Although these data have sometimes led to conflicting conclusions (e.g., Mosteller and Moynihan, 1972; Jencks et al., 1972), major reanalyses have tended to support the basic findings of the Coleman report. Mayeske and his associates at the U. S. Office of Education performed a series of reanalyses of the EEOS data, using different statistical techniques and going deeper into the relationship between the school and the student. In the first of these analyses (Mayeske, Wisler, Beaton, Weinfeld, Cohen, Okada, Proshek and Tabler, 1972), they sought to distinguish the influence of the school's characteristics from all other kinds of influence, after equating schools for differences in size and socioeconomic characteristics. Their conclusions were that the children who benefited most from their schooling were white or Oriental-American and came from well-to-do homes in which both parents were present. They also found that schools performing well on one educational outcome tended to perform well on other outcomes, and that the most influential school variables were those connected with teacher characteristics.

In the second analysis (Mayeske, Okada, Beaton, Cohen and Wisler, 1973), individual student achievement was correlated with such variables as home background, community, ethnicity, socioeconomic status (SES) and attitudes toward education. This study found that educationally related child-rearing practices play a greater role in a student's achievement than either his family's SES or stability; that family background factors play an independent role in achievement nearly five times that of school factors; and that the role of the total student body's values and achievement level was six times that of any other school factor.

The third analysis (Mayeske, Okada and Beaton, 1973) used student attitudes rather than achievement as the criterion and investigated the relation of various school, family and student factors to these attitudes. This study concentrated on a set of variables from the EEOS questionnaires called "Attitude toward Life"; these variables were thought to measure the degree to which a student believed he could control his own life. This belief was found to be strongest among white students and to be influenced more by the characteristics of the student's peers than by his teachers' characteristics or even his own school performance. One of the conclusions reached was that this important attitudinal variable might be improved for all students by altering the performance criteria and reward structure of schools.

A number of other studies have also pursued the question of whether or not schools do indeed make a difference and, if they do, what are the critical factors. The relation between the amount of schooling and the amount of achievement was explored by Wiley (1973), who hypothesized that the effects of various components of schooling will vary with the student's exposure to them. The analysis used EEOS data for the sixth grade sample from the Detroit metropolitan area. Outcome variables included verbal ability, reading comprehension, and math achievement. School (input) variables were average daily attendance, number of hours a day, and number of days a year devoted to different subject matter areas. Such individual background variables as ethnic group, number of children in family and number of possessions in home were also used. Wiley found, first of all, that typical students in some schools receive 50% more schooling than students in other schools. Further, he found that an increase in the number of hours and days devoted to instruction lead to significant gains in achievement. He concluded that quantity of schooling is an important determinant of achievement.

Klitgaard and Hall (1973) used data from several large-scale surveys to locate exceptional ("overachieving" or "outlier") schools. For their study they first controlled for nonschool factors such as SES, minority enrollment, regional variables and previous test scores. All residual variation was assumed to represent the school effects. Results were mixed and not at all clear-cut. Generally, they found that distributions of school achievement scores were always extremely "tight," after nonschool background factors were held constant--that is, no immediate evidence of extreme overachievers was found. They did find some unusually effective schools and districts, but no single school had raised student achievement scores more than about 0.8 of a standard deviation (inter-student). Unfortunately, this study did not include information about school programs, so it is not possible to say why some of the schools and districts seemed to be more effective than others.

A study of the impact of various compensatory reading programs on student achievement is now being carried out in 254 schools throughout the country (Rubin, Trismen, Wilder and Yates, 1973), based on descriptions of program characteristics provided by teachers and principals. The final analysis is expected to provide an indication of the possible effects of different approaches on student achievement in reading and on student attitudes toward reading and toward themselves as readers. Since one of the variables in this study is the amount of class time devoted to reading instruction, it is hoped that it will provide further evidence of the relative importance of amount of schooling.

A somewhat different perspective on the issue of school/student interaction was provided by the Study of Academic Growth and Prediction, which followed a nationwide sample of students from 1961 to 1969 (Hilton, Beaton and Bower, 1971). In this analysis, groups were classified according to sex, ethnic group, high school curriculum and father's education; and the stability of test scores and certain scales from the questionnaire were examined over time. It was found that relative stability was highest for academic ability, next highest for information, and lowest for interest and extracurricular activity scales. This suggests that schools might have greatest impact in attitudinal areas if these are the focus of attention.

For the past several years, general outcomes of schooling have been the focus of study for the National Assessment of Educational Progress (Womer, 1970), which uses data collected from different age groups in different sections of the country. The data analyzed by NAEP to date reflect group differences in a number of subject matter areas. However, the major objective has been to establish nationwide baseline data for school outcomes and the study does not extend to the definition or measurement of input and process variables.

In spite of some intriguing clues that significant school effects may yet be found by pursuing new paths of inquiry, the evidence to date consistently indicates that student background characteristics are of primary importance to achievement. This conclusion is not peculiar to the United States. In England, the Plowden study carried out in 1964 demonstrated the importance of background and early learning factors in that country (Central Advisory Council for Education, 1967). A follow-up survey of children in the Plowden sample reiterated this finding (Peaker, 1971).

On a larger scale, a series of reports from the International Project for the Evaluation of Educational Achievement (IEA) involving a worldwide sample of students in both English- and non-English-speaking countries, also supported the primacy of student background characteristics as predictors of achievement. For example, the study of achievement in mathematics in 12 countries (Husen, 1967) found that school variables gave generally small and inconsistent relationships to achievement; but a significant relationship was found between parents' characteristics and math achievement. The study of literature education in ten countries (Purves, 1973) found that, although there was little consistency in school variables among countries, home background nevertheless accounted for a great deal of the variance among schools. The study of science education in 19 countries (Comber and Keeves, 1973) found that home background contributed considerably to differences in achievement, while variables associated with the program and type of school accounted for moderate proportions of total variance, and learning condition variables contributed less to the total variance than either home or school variables.

More recent reanalyses of a portion of the IEA data by Coleman, however, suggests "that reading achievement is more fully an outgrowth of home influences than are either of the other two subjects [science and literature], less a function of what takes place at school" (1975, p. 382). Coleman also noted that school effects (relative to the effect of the home) have probably been underestimated because of the common use of tests requiring reading skills.

When interpreting the general uniformity of findings from prior research into school effects and student achievement, it must be emphasized that most large-scale studies have concentrated on school-resource effects and have not examined the impact of specific educational practices. Four excellent reviews of the literature, which summarize the limitations present to some degree in nearly all previous studies, have recently been published (Spady, 1973; Averch, Carroll, Donaldson, Kiesling and Pincus, 1972; Bryant, Glaser, Hansen and Kirsch, 1974; Michigan State Department of Education, 1971). These limitations fall into four general categories: (1) unsophisticated measurement and inadequate statistical methods; (2) lack of longitudinal data; (3) lack of individual or classroom-level data; and (4) measurement of outcome only in terms of achievement, with little or no attention to other kinds of school outcomes such as student attitudes. These issues were all considered in the design and implementation of Project LONGSTEP. By its attention to such concerns in the early 1970s, the study represents one of the first large-scale applications of the input-process-output approach to the investigation of educational effects.

Studies of School Processes

The teacher is a very important resource in any classroom, and thus the major part of previous research on school processes has focused on the role and the characteristics of the teacher. This research has been undertaken from two distinct perspectives: (1) an emphasis on personal teacher traits such as intelligence, attitude toward teaching and an understanding of pupil needs; and (2) an emphasis on resource-related factors such as salary, tenure, training and experience. Attempts to interrelate the findings of these two general approaches have so far proven unsuccessful (Glass, 1974).

Rosenshine (1970) reviewed efforts to correlate measures of teacher quality with residualized student achievement gain scores, but he found no consistent results. He pointed out that in nearly all studies of this type, classrooms from different schools are pooled together in the same analysis, thereby creating considerable confounding due to the interaction of neighborhood characteristics, school differences and home environments. Flanders (1970) argued that the comparison of teachers according to the performance of their students can be misleading; the measures of teacher quality that he recommends are based on extensive classroom observation and interviewing of both students and teachers.

Attempts to associate the more easily quantified teacher variables such as age, units of credit, degrees, salary and tenure, with student performance have also met with little success. Formal credentialing has become so prevalent that its impact within the teaching profession has been virtually eliminated (Hanushek, 1970). Nearly every study on the effects of teacher experience presumes a linear relationship between length of service and skill or attitude, even though there is no real empirical evidence for its validity (Spady, 1973). Although there are some indications that the overall power of districts to offer higher salaries (i.e., they have more money) is associated with student performance, there is no consistent evidence that this is due to higher teacher quality.

In short, despite years of research there is little empirical evidence to date that identifies desirable teacher characteristics or that explains how these characteristics might be fostered by training and measured in the classroom, and how they relate to student outcomes.

It has been recommended that the only way to define "good" teaching is by observing teachers in action; consequently, growing attention has been paid over the past few years to the interactions in the classroom. A rich body of literature describes new methodologies and strategies for observing and coding the verbal and nonverbal behaviors of students and teachers (see Simon and Boyer, 1967 ; Boyer, Simon and Karafin, 1973). However, little of this research attempts to relate such behaviors to pupil outcomes.

Another deficiency in the observational literature is the lack of any unifying and generally accepted assumptions concerning which classroom processes it is important to study. Nearly all systems reflect a bias on the part of the individual researcher, who has developed a method, strategy or technique to observe some limited characteristics (whether verbal or nonverbal interactions, psychological climate, or management of materials) thought to have an effect on student achievement. Very few of the systems have been designed to describe the total classroom environment as a basis for developing hypotheses about which are effective teaching-learning behaviors.

One exception is an observational system recently developed for evaluation of the various teaching models encompassed by the Follow-Through Program, ranging from highly structured to highly individualized approaches (Stallings, 1974). In this study, 28 classroom process variables were recorded and then correlated with such student behaviors as independence, task persistence, cooperation, and question asking, as well as test scores in reading and math. An analysis of data from 105 first-grade and 58 third-grade classrooms showed significant correlations between high test scores in reading and math and small-group instruction for first grade; large-group instruction for third grade; stimulus-response-feedback interactions between the student and teacher; self-instruction and task persistence; and time spent in reading or math activity. In general, it was found that highly structured classrooms had the highest scores in reading and math, while more flexible classrooms tended to produce low absence rates, high independence, and high scores on a test of nonverbal, perceptual problem-solving. The author concluded that classroom process variables are more predictive of child behaviors, absence rates and test scores than are entering-school test scores for students in grade one and grade three.

The traditional, heterogeneous classroom has been strongly criticized in recent years for its failure to provide for individual student differences in learning needs and abilities. As a result, there has been a major movement in the schools toward the use of more individualized instructional programs and more "open" classroom structures. However, to date there has been little evaluation of the comparative effectiveness of the traditional versus the innovative or individualized approach. The literature in this area tends to be theoretical and descriptive, rather than analytical or evaluative (Weisgerber, 1971a and 1971b; Schramm, 1964; Walberg and Thomas, 1971; Spådek, 1970; Hertzberg and Stone, 1971); and there is a lack of reliable information on the impact of innovative materials and practices on student outcomes. A recent study conducted by the staff of the University Elementary School of UCLA (Goodlad, 1970) outlines one possible reason why such explorations have not been made. Specifically, this study was designed to find out whether some of the more innovative educational practices recommended in recent years were actually finding their way into the schools. More than 250 classrooms in a nationwide sample of districts were visited by Goodlad and his associates. In general, observers found labels such as "team teaching" applied to a wide variety of practices. Furthermore, in none of the schools was there a comprehensive plan to integrate innovative practices throughout the entire school program. The research staff concluded that "some of the highly recommended and publicized innovations of the past decade were dimly conceived and, at best, partially implemented in the schools claiming them" (p. 72).

In light of the problems encountered by researchers in defining and identifying differences among school programs, it is hardly surprising that there has been little research done on the effects of various educational processes or program "treatments" on the individual student. Those actively concerned with school programs, both innovative and traditional, have called repeatedly for a systematic and long-term study not only of input-output variables but also of the effects of educational processes over a period of time. As Dyer (1972) has pointed out:

> Rational decisions on long-term educational policy must rest on nothing less than long-term studies of educational practices which follow samples of the same children through large segments of their school experience--say four or five years. For only so is it possible to get a reasonably dependable check on the relative importance of the innumerable factors that are in fact shaping pupil development (p. 406).

Project LONGSTEP, the Longitudinal Study of Educational Practices, was designed to be such a long-term study. It incorporates measures of the educational process and measures of student input and output to investigate the impact of intensive educational innovation. Project LONGSTEP was one of the first large-scale studies to utilize this approach.

II. TECHNICAL APPROACH

A. Introduction

In February 1969 AIR staff members began preparing the research methodology that would be required for a large-scale study of the effects of various educational approaches on a number of different student populations over an extended period of time. An underlying rationale was that cognitive and affective variables are complex measures whose variance can be related to a number of sources. Some of these sources, such as socioeconomic status (SES), are typically considered beyond the influence of the school and its programs and teachers. Other sources, such as program characteristics, school characteristics and even teacher characteristics, are more amenable to control by the school. It was decided, therefore, that this study would attempt to assess the effects that such controllable factors have on student performance and attitudes, over and above the effects that uncontrollable influences have. The study consequently was designed around the following three basic areas of interest:

1. Educational Treatments. In the broadest sense, educational treatments were considered the total educational environment into which a given student entered when participating in a school program. Measures in this category represented various aspects of each student's educational experiences, including specific program processes and certain teacher characteristics.

2. Student Characteristics. These variables included the characteristics students brought with them each year into the educational environment provided by the school. Included here were background variables such as SES, age, sex, cultural or ethnic group, as well as attitudes and achievements from prior years.

3. Student Outcomes. Outcome measures included student cognitive achievement as measured by standardized achievement tests, and attitudes as measured by certain items in the student questionnaires. In the terminology of systems analysis, such variables are considered output variables; they provide the criteria for the evaluation of educational impact.

In the fall and spring of the 1969-70 school year, the data collection procedures and instruments were field tested in a number of school districts across the country. (Two of these districts would later participate in the full-scale implementation of Project LONGSTEP.) The field tests indicated that the study plan was feasible, and in the summer of 1970 AIR received approval to implement Project LONGSTEP on a long-term basis.

The remainder of this section describes in detail the selection of the final sample of schools and students to be included in the study, the development of the data collection instruments and procedures, and the development of the data management and quality control system.

B. Selection of Programs, Schools and Students

Program Selection

Project LONGSTEP was designed to determine, across a projected three- to five-year period of time, if intensive innovative educational efforts were related to substantial differences in student outcomes. Because four of the five sites included in the design phase of the study, at the direction of USOE, were sites emphasizing dropout prevention, the total sample was not sufficiently diverse to enable examination of the study's key hypotheses. USOE and AIR both recognized that, in order to achieve the goals of the study, a much broader range of educational strategies would have to be sampled.

Criteria for program selection could not be statistical in nature because the universe of verified large-scale innovative efforts was not known. Furthermore, since only a limited number of sites would be able to participate in Project LONGSTEP, a randomly selected sample would not be likely to contain the diversity or the intensive level of innovative effort necessary to achieve the study objectives. Accordingly, a nationwide search for programs incorporating intensive innovative practices was undertaken.

A first step in this process was the establishment of a detailed set of guidelines for identifying programs worthy of consideration as candidates for the study. Guidelines for program selection emphasized program scope and intensity, instructional content, anticipated program continuation, and expected cooperation in a longitudinal study.

Guidelines pertaining to program scope and intensity reflected the decision to include in the study innovative educational strategies representing a sizeable investment of school resources and impacting on a significant proportion of students at different grade levels.

Guidelines in the instructional content area emphasized specificity in strategies and outcomes, variety of treatments within programs, and a clear rationale for departing from traditional class instruction. A further requirement in program selection was a high probability that the program would continue in operation for at least three years. On very practical grounds, programs that were already the subject of intensive study by other organizations were not to be considered as candidates in order to avoid further burdening of local staff and students. Another important selection requirement in a study involving voluntary participation over a three- to five-year period was some reasonable assurance of continued cooperation on the part of district and school administrators, teachers and students.

The complete set of guidelines for the selection of candidate programs is contained in Appendix II-A.

A multistage screening process, consisting of literature search, direct program contact, outside consultant review, and site visiting was employed for program selection.

Literature search. The first step in program selection consisted of reviewing available literature. A primary source book for this effort was Pacesetters in Innovation (U. S. Office of Education, 1969), a compendium of abstracts describing all Elementary and Secondary Education Act (ESEA) Title III programs in operation as of February 1969. Documents collected in conjunction with previous AIR educational evaluation projects provided valuable additional sources. Leads to still other programs were obtained through personal contacts and miscellaneous reports and newsletters.

Information about several thousand programs was screened, with approximately 150 programs "surviving." A second screening, in which the same criteria were applied but with somewhat greater rigor, was then conducted. Approximately 50 of the originally selected programs survived this screening.

Direct program contact. Cognizant personnel affiliated with each of the 50 programs that remained after the initial screening were contacted directly by telephone. Of primary concern at this time was the likelihood of program continuation for the duration of the study and the accuracy of the descriptive information used for initial program screening.

Only eight of the fifty programs surviving the literature search remained eligible for consideration after first-hand information was obtained. Some programs had terminated when Title III funding was withdrawn; some did not appear likely to continue for the next three years; but the most common cause for elimination was that programs did not measure up to their early publicity (i.e., the programs as actually implemented were found to differ substantially from their proposal abstracts in the Pacesetters directory).

Outside consultant review. In order to provide an independent check on the results of our program selection process, the services of outside consultants were obtained. Brief descriptions of the eight programs that had been identified to this point were mailed to seven consultants, [illegible] nationally known authorities in education. They were asked to review the descriptions, comment on any of the programs with which they were familiar, and nominate additional candidates that had been overlooked by the project team. For the most part, comments about the eight programs identified by AIR were positive. In addition, 23 programs that had not already been identified by AIR were nominated by the five consultants responding. Project staff contacted these programs by telephone and used the same screening strategy employed with the 50 programs found during the AIR literature review. At the conclusion of this step, final decisions were made concerning which of the programs identified to date best met the study criteria and merited site visiting.

Site visiting. The purposes of the site visits were to obtain more complete information on the content of the educational program and its context, and to acquaint the school district representatives with the benefits and requirements associated with participation in the study. Teams of USOE and AIR representatives visited a total of 20 candidate school districts.

Following the visits, numerous telephone and mail contacts were made with various school representatives in response to questions that arose after the site visits. AIR staff then prepared a list of tentative recommendations for program selection based on the results of the site visits and on the study of the descriptive information obtained on candidate programs. The recommendations were then discussed with the USOE Project Officer, and a final list of candidate programs was agreed upon.

Factors above and beyond the criteria established earlier were taken into account in the selection of the final set of programs to be included in the study. This was necessary in order to provide a sufficient range of programs and population characteristics. Specifically, the following factors were considered: diversity of program content, grade range included, average income and range of incomes of students' families, per-pupil expenditures, percentage of high school graduates who go on to college, and extent of minority group representation. Formal invitations were then extended to the school districts selected. By this time, AIR staff had sufficient contact with the school districts to be reasonably certain of their willingness to participate.

A disquieting result of this intensive search for sites, based on literature review and consultant recommendations, was the under-representation of large cities having a concentration of inner-city schools.

None of the major metropolitan areas in the Northeast, East Coast or Midwest sections of the country was represented on the candidate list. Only two cities with populations of more than 500,000 were initially identified, one in the Pacific Northwest and one in the South. There appeared to be three primary reasons for this lack of inner-city schools in the original sample: (1) the programs that had been identified in such schools failed to measure up to their early publicity; (2) programs that may have met the study criteria had been dropped or were facing probable termination at the end of Title III funding; and (3) many inner-city programs were experiencing implementation problems, including lack of stability and uncertain duration, which were apparent to AIR staff members and/or to the consultants involved.

Both AIR and USOE personnel felt that it would benefit the study to include more programs involving low SES inner-city students, providing such programs met the selection criteria. Therefore, a second search, focusing on large-city schools, was conducted early in the 1970-71 school year. During this activity, project staff members again reviewed recent literature and findings from previous and ongoing AIR evaluation studies. The nomination of candidate programs was also again solicited from AIR staff and outside consultants.

As a result of this second search, seven additional programs were tentatively identified as candidates. Further study showed that six of these programs did not meet the criteria for inclusion in the study. Primary reasons for dropping these six programs from consideration were: limitations on the number of schools or classrooms involved; emphasis on remediation rather than innovation; emphasis on preschool children; restrictions on subject matter coverage; and finally, indications of internal friction and possible lack of support for the research study.

The single remaining program was then contacted by telephone and detailed information on the study was mailed to the district. In reply, the school district representative described several problems that would make this district's participation difficult, both for local schools and for the study. These problems involved violent student resistance to additional testing, resistance to group testing on the part of the teaching staff, and possible teacher dissatisfaction even if paid for extra work outside the classroom. Negotiations were therefore discontinued, and it was decided that a further search for large-city districts would not be possible within the budget and schedule constraints of the study.

Later experiences during the implementation phase of the study indicated that the problems enumerated by this district were not unique, but rather were becoming more common among large-city school districts. In addition, such factors as high transiency among students, high teacher turnover, student unrest due to racial and ethnic conflict, and political issues such as busing also made the successful conduct of longitudinal research in inner-city schools extremely problematical.

Altogether, 14 school districts agreed to participate in the implementation phase. Three of the locations had participated in the initial design phase of the study; the remaining 11 locations were added during the 1970-71 school year. During the 1971-72 school year, one of the school districts was dropped from the study by mutual consent because the program of interest was being shifted to a different group of schools, and test scheduling conflicts and other administrative difficulties would have resulted. The 13 school districts that did participate throughout the implementation phase of the study were located in nine states: California, Florida, Kentucky, Minnesota, Pennsylvania, Texas, Utah, Washington and Wisconsin.

Description of Participating Programs

The 13 sites included in the study provided variability along a number of dimensions. For example, according to information obtained during the 1970-71 school year (as shown in Table II-1), the communities serviced by the school districts ranged in size from 2,500 to over 600,000 and varied from rural to urban-metropolitan in setting. The diversity in socioeconomic level was notable, as evidenced by the percentage of students in the participating schools receiving free or reduced-price lunches. This ranged from less than 1% to nearly 30%. The reported instructional costs per pupil in the participating districts

TABLE II-1

Participating School/Community Characteristics Based on Data Obtained during 1970-71 School Year

Location[1]	Community Size[2]	Community Setting[3]	Cost per Pupil[4]	No. of Students (1970-71)[5]	Percentage - Minority[6]	Percentage - Lunch Aid[7]
Site A	31,100	Urban - non-metropolitan	$1,050	1,820	33%	28%
Site B	524,200	Urban - metropolitan	$ 820	2,050	29%	21%
Site C	49,400	Suburban	$ 890	940	< 1%	< 1%
Site D	4,000	Rural	$ 680	120	< 1%	10%
Site E	14,900	Rural	$ 540	1,780	1%	15%
Site F	14,200	Suburban	$1,010	1,260	4%	3%
Site G	100,600	Urban - metropolitan	$ 790	3,790	7%	20%
Site H	4,000	Rural	$ 760	140	2%	19%
Site I	620,100	Urban - metropolitan/ suburban	$ 800	3,730	12%	< 1%
Site J	250,000	Rural/ suburban	$ 620	1,990	4%	12%
Site K	5,000	Rural	$ 680	370	< 1%	8%
Site L	2,500	Rural	$ 760	90	3%	22%
Site M	65,000	Urban - non-metropolitan	$ 640	1,550	21%	2%

Footnotes to Table II-1

[1] To preserve site anonymity, the order of sites in this table does not correspond with the order shown in later report tables or with that used in the study's data base.

[2] Community size is an approximation of the population living in the school district service area.

[3] Operational definitions used in this study are the following:

Urban - metropolitan is a community of high population density, with such characteristics as a downtown area and ghetto sections. It has near it at least one other urban center, and together these form a metropolitan area.

Urban - nonmetropolitan is a community with high population density which may not have a highly developed downtown area or be close to another urban center.

Suburban is a community with moderate population density on the periphery of a city. It is often a bedroom community and is frequently without a ghetto area.

Rural is a community with low population density, which is generally not close to a large urban center and is often agricultural in character.

[4] Cost per pupil is an estimate of instructional costs, rounded to the nearest $10. Strictly speaking, these figures cannot be directly compared with each other since they are not adjusted to reflect cost of living in the various communities. Also, the method by which cost per pupil is calculated varies somewhat across sites and may be based on slightly different time periods, e.g., 1970 or 1971.

[5] The figures shown are the approximate number of students in the sample during 1970-71, based on spring enrollment estimates by the school districts for the grade levels and classes in the study.

[6] The minority percentage shown is a composite of all minority groups. It is the percentage of minority students enrolled in the sample schools and grades, using school approximations.

[7] The term "lunch aid" includes both free and reduced-price lunches. The figure shown is the percentage of students receiving lunch aid in the sample schools and grades, using school approximations.

varied from a low of $540 to a high of $1,050. A final indicator of the diversity in the sample was the percentage of minority students in the participating schools, which ranged from less than 1% to over 30%.

The participating schools in the 13 school districts spanned a wide range of educational practices, from the very orthodox conventional classroom to a highly diversified collection of innovations. Variation in educational practices occurred both within and across the participating districts. An overview of the major educational features present in our sample at the time programs were selected for inclusion in the study, and the distribution of such features across the sites, can be seen in Table II-2. Caution should be exercised in drawing conclusions from this table, however, since the same label may have encompassed significantly different activity elements from one school district to another (Goodlad, 1970). By the same token, different labels were often applied to very similar activities. Further, it should be stressed that the educational activities listed in the table were often given these designations by the school systems. They did not necessarily reflect the underlying educational processes that were the primary concern of this study. Measures of educational processes used in Project LONGSTEP are discussed in a later section. Nevertheless, Table II-2 does provide a gross indication of the wide variation that existed in our sample.

Selection of Schools and Students

While the study spans Grades 1 through 12, not all schools and grades in each district were included in the sample. Final decisions as to the schools and students who would be in the study were based on several considerations. First, schools were selected which best met the criteria of program scope and intensity, innovative instructional content, anticipated program continuation and expected cooperation in the study. As noted previously, Appendix II-A provides a detailed description of these criteria. In addition, an effort was made to include a number of more conventional schools in order to ensure adequate diversity in treatment variables.

TABLE II-2

Overview of Major Educational Features Occurring in One or More Schools of the District Participating in Project LONGSTEP

Major Educational Features	School Districts*												
	1	2	3	4	5	6	7	8	9	10	11	12	13
Small-group activities	X	X	X	X				X	X	X		X	X
Team teaching	X	X	X	X	X	X	X	X	X				
Multimedia emphasis	X	X	X	X	X	X	X	X	X	X		X	
Ungraded curriculum		X	X	X	X		X	X	X				
Large-group instruction		X	X					X	X	X	X	X	X
Independent study	X	X	X	X	X			X		X		X	X
Unique school design		X	X				X	X					
Differentiated staffing	X	X				X							
Volunteer aides	X	X	X	X	X	X	X	X	X	X			
IPI reading/math		X			X	X			X				
Project PLAN							X		X				
Individualized instruction (other than PLAN or IPI)	X	X		X		X	X	X	X	X		X	X
Student selection of materials	X	X	X			X		X	X	X			
Modular scheduling	X	X	X	X				X					
Teacher sensitivity training						X							
Intensive guidance and counseling						X							
Home visitation						X							
Traditional/conventional classrooms	X		X	X	X	X	X	X	X	X	X	X	X
Teacher developed curriculum materials	X	X	X	X		X		X	X	X	X	X	X

*To preserve site anonymity, the order of sites in this table does not correspond with the order shown in later report tables or with that used in the study's data base.

Second, given the hypothesis that intensive innovation would have substantial impact on student outcomes, it was desirable that, wherever possible, the sample should include students who had not yet been exposed to these intensive practices. For this reason, we attempted to select for first-year participation students in the grade below that at which the major innovations were targeted. One instance where this was not possible was when the program included first graders.

A third consideration in the final determination of the school and student sample within each participating district was the configuration of the normal stream of feeder schools to higher grade level schools. Because of LONGSTEP's longitudinal strategy, students tested the first year were to be followed each subsequent year of the study for as long as they remained in that school or in another participating school. In order to minimize the dilution of the sample over time, schools in which the majority of the students would be lost to the study when they moved into nonparticipating schools were not selected.

Lastly, with the exception of special education students, all students in a grade within each participating school were included in the study, whether or not they were involved in an innovative program. This added diversity to our sample and enabled us to provide schools with test scores for complete grades. The furnishing of such data was, in fact, the primary incentive offered to the school districts for participation in the study.

The special education students who were not to be included were those with IQs of 80 or below or those who were more than three years behind their age peers. Such students were not included because (1) it was expected that their response to educational treatment would be radically different from that of the majority of students; (2) in many cases, while they might have been assigned to a participating grade for administrative or social reasons, they often were given very different treatments in the areas of language arts and mathematics; and (3) teachers felt that such testing could result in serious emotional problems for these students if they attempted to complete the level of the test scheduled by this study.

As the reader will note from this discussion, the study did not involve a random sample of students or a random sample of school programs (or even a sample of school programs to which students could be randomly assigned). By necessity, the study was designed to be a nonexperimental investigation since the rudiments of experimental control simply could not be implemented in schools that were voluntary participants. The control of the treatments in the study rested, in fact, in the hands of the communities, school boards, teachers and principals who were attempting to provide the best education they could for their children. Their concern was with effective schools and not with the random assignment of students to particular programs, with the systematic variations of program components, or with other considerations necessary for rigorous statistical inferences. As a consequence, Project LONGSTEP was specifically designed to examine the correlates of student output demonstrated in naturally occurring educational groups like those participating in the study.

Although we could not randomly select students or assign them randomly to the various educational treatments being studied, programs were purposefully selected so that the total sample of students would represent a wide range of groups. And, as mentioned previously, all students in the grades selected in a given school were included in the study, not just those involved in a particular educational program. These selection rules ensured diversity, both in student and in treatment characreristics, and helped to avoid possible bias due to sampling selected classrooms within schools.

C. Data Collection Instruments and Procedures

The study design required the collection of a great deal of different kinds of information on students and their educational environment. It was felt that such data would not only lead to a greater understanding of the processes involved in educational innovations but would also allow the evaluation of possible alternative explanations for any treatment effects that might be observed.

The data collection instruments used in the study provided information on student cognitive performance, student characteristics, student attitudes, teacher characteristics and educational treatments. Student cognitive achievement was measured by standard, commercially developed instruments. Student characteristics and attitudes and teacher characteristics were assessed by questionnaires developed specifically for Project LONGSTEP. Educational treatment characteristics were documented by specially designed instruments developed for staff use during site visits. Copies of the final questionnaires and treatment documentation instruments will be deposited in the Educational Resources Information Center (ERIC) at the conclusion of the study.

Student Background and Outcome Measures

Cognitive instruments. The selection of a standardized achievement test that was sensitive to changes in basic skills but was not dependent on specific curriculum content was essential (1) because the goal of Project LONGSTEP was to examine the impact of intensive innovation on student achievement across a sample of programs selected partially on the basis of their diversity; (2) because programs that varied in terms of educational strategies could be expected to differ with respect to specific instructional objectives and content; and (3) so that the achievement gains of students participating in Project LONGSTEP's select sample could be compared with national norms. Prospective tests were evaluated according to the following criteria:

(1) Acceptability in terms of standard psychological measurement criteria related to item characteristics, standardization procedures, and overall reliability and validity measures.

(2) Ease of administration, processing and scoring, and other administrative and logistics considerations.

(3) The degree to which the test was considered relevant and interesting to present-day students.

(4) Attractiveness of format and acceptability to the students.

(5) The extent to which the instrument was independent of any particular content material or curriculum orientation. (An instrument oriented to a specific curriculum would certainly be "unfair" to some programs.)

(6) The absence of specific grade level information on the test booklet. This would enable use of the test in both ungraded and graded classrooms without causing concern on the part of the student that he was being tested above or below his grade level.

Initial searches for instruments identified seven appropriate standardized achievement batteries available at the time (1969). Each of these was examined in light of the stated criteria. In addition, the tests were discussed with specialists knowledgeable of the latest developments in achievement testing.

The Comprehensive Tests of Basic Skills (CTBS), published by CTB/McGraw-Hill in 1968 and standardized on over 200,000 students from public and private schools in all 50 states, was selected as the battery that best met the selection criteria for use in Grades 3 through 12. It contained up-to-date content and was designed to reflect a balanced sampling of items from the various curricular approaches used in American schools rather than being restricted to a specific curriculum orientation. The major areas measured by the test were reading, language, arithmetic, and study skills. The number of test booklets required for administration was minimal and the directions for administration were explicit and straightforward--essential features for a test to be administered by school personnel in large-scale testing situations. Futhermore, the absence of a grade/level designation on the test booklet made it attractive for testing in mixed grade or ungraded situations. Also, an examination of the technical publications provided for the battery indicated that it was acceptable in terms of standard psychological measurement criteria. Reliability estimation methods included an assessment of the internal consistency, interlevel articulation (i.e., the extent to which different levels of the test measure the same things), interform or parallel form reliability, and interlevel/interform reliability. All methods resulted in moderate to high reliability estimates for the Reading, Language and Arithmetic Total subtest scores.

The CTBS had four levels, each appropriate at certain grade levels:

Level	Grades
1	2.5 - 4.9
2	4.0 - 6.9
3	6.0 - 8.9
4	8.0 - 12.9

The overlapping of grade levels provided a degree of flexibility that was another desirable characteristic of the CTBS.

The test battery yielded a variety of meaningful scores. In addition to percentile ranks, stanines and grade equivalent scores, CTBS raw scores could also be transformed into "Expanded Standard Scores" (CTB/McGraw-Hill, 1970) or Expanded Scale Scores. These Expanded Scale Scores enabled a student's growth to be charted from grade to grade regardless of which level or form of the CTBS was taken.

The selection of instruments for use in Grades 1 and 2 was more difficult. Although we would have preferred to use one cognitive battery for all grades participating in the study, the CTBS was inappropriate for fall testing in the second grade, and entirely inappropriate for the first grade. Considered in the selection of the instrument for Grades 1 and 2 was the problem associated with the use of reading and math tests at these early primary grades. Because of possible wide differences in the time at which formal reading and math programs were introduced in the participating districts, using tests that measured specific reading and math skills would present the problem of some sites and students being unduly penalized. Nevertheless, some type of baseline measurement was desired for participating first and second grade students. A group intelligence test designed to tap prerequisite competencies related to subsequent academic achievement appeared to be the most appropriate type of instrument.

The California Short Form Test of Mental Maturity (CTMM), 1963 Revision, published by CTB/McGraw-Hill was selected. An examination of the instrument and various published reviews indicated that it would provide appropriate baseline measures for students in Grades 1 and 2[1]. Seven scores were provided: logical reasoning, numerical reasoning, verbal comprehension, memory, language total, nonlanguage total, and overall total.

Typically, the tests were administered by classroom teachers or other school or district staff. Although project funds did not permit the hiring of special test administrators, AIR did provide checklists and other instructional materials. Also, in nearly all cases AIR staff were present in the district during testing in order to answer questions and provide other assistance. The participating districts were as motivated to obtain accurate test results as were AIR staff since they had been promised the test results for their own use. Consequently, special efforts were often made to obtain as good a level of test conditions as possible in line with the suggestions of the test producer. For example, the testing of students in Grades 1 and 2 was typically conducted in small groups.

Student questionnaires. Two primary considerations led to the inclusion of student questionnaires in the study: (1) the desire to measure noncognitive or affective, as well as cognitive, effects of the educational environment; and (2) the desire to gather data relevant to certain biographical, socioeconomic, racial and ethnic characteristics of the students involved.

Attitudinal variables were intended to serve two functions in this study. First, they could be included as independent variables, with cognitive achievement as the dependent variable. Second, since selected attitude changes also constitute very real and educationally significant effects that educational treatments may have on the student, such measures could be viewed as dependent variables as well.

Various strategies for measuring student attitudes were investigated during the design phase. Such techniques as direct observation, the

---

[1] Beginning with the 1971-72 school year, student testing occurred only in the spring, as will be discussed later. Consequently, for this year and the following year, second grade students were tested with Level 1 of the CTBS.

Critical Incident Technique (Flanagan, 1954), or similar behavioral measures would have provided the most objective indices. The scope of the study, however, precluded the possibility of obtaining this type of information for every student. Peer group ratings were eliminated from consideration for similar reasons and because of difficulties in obtaining comparable information across the wide span of student ages. Questionnaires, despite their many inadequacies, were the only feasible approach to measuring student attitudes within the constraints of the study.

It is generally accepted that students with favorable attitudes toward themselves and their teachers will do better in school and, as noted previously, recent research (Coleman et al., 1966; Mayeske, Okada and Beaton, 1973) has shown a relationship between achievement and a feeling of control over one's life. While research results on student persistence, motivation and self-concept are not consistent, there are indications that attitudinal areas of importance include (1) self-concept/ personal appraisal, (2) attitude toward education in general, (3) attitude toward school, (4) attitude toward teachers, (5) attitude toward fellow students, and (6) attitude toward society. The questionnaires were designed to cover these six attitudinal areas.

As far as possible, items were based on existing sources such as Project TALENT (Flanagan, Dailey, Shaycoft, Gorham, Orr and Goldberg, 1962) and the Equality of Educational Opportunity Survey (EEOS) (Coleman et al., 1966). The decision to use existing items whenever possible was based on the assumption that extensive effort had already been devoted to the development, tryout and revision of these items. It was also considered important to be able to compare the results of this study with those of previous large-scale studies. Revisions were made of some items, as, for example, when data from the Coleman report indicated that some of the choices to a particular item had been chosen by a very small proportion of the population or where the "I don't know" choice was selected by a majority of respondents.

Where the pool of existing items did not appear to provide adequate coverage of a particular attitude area, new items were written by members of the project staff. Also, because different versions of the questionnaire were required for students at different grade levels, some items were modified to allow for variations in reading ability and for applicability to a particular grade level. For example, questions about post-high school plans were asked only on the questionnaire for Grades 9 through 12.

In addition to attitudes, student questionnaires covered such background factors as amount of preschool training, the socioeconomic status of the students' parents, and family structure and stability. Although much of this information was available in school records, the size of the sample made it necessary to obtain such data from the questionnaires. Once again, items were drawn from existing sources and modified for appropriateness. In some cases, items that were used in other studies but that had proved of little value were rejected. In examining the EEOS report, for example, it became obvious that an almost equally reliable index of socioeconomic status could be obtained with far fewer items than had been included in that study.

Five student questionnaires were developed: Form A, to be completed by the teacher, for Grades 1 and 2; Form B, to be completed by the teacher, and Form C, to be completed by the student, for Grades 3 through 5; Form D, to be completed by the student, for Grades 6 through 8; and Form E, to be completed by the student, for Grades 9 through 12. Insofar as possible, the same items were used across all forms, with only minor changes in vocabulary, and with the addition of items about student attitudes toward higher education and occupations on Forms D and E. For the lower grades, student attributes were reported by teachers on the basis of such observations as the child's ability to make friends easily and the degree of confidence exhibited by the child in his relationships with adults.

Teacher Questionnaires

Any study of educational effects must take into account the influence of the professional and personal characteristics of teachers on the achievement, attitudes, or other school-related behaviors of their students. It was therefore deemed important in the present study to collect information concerning the background, behavior and attitudes of teachers. The levels of training and experience of the instructional staff, the extent of their commitment to the education profession, their empathy for the problems of students, and their attitudes toward their undergraduate education and present working environment were the characteristics selected as most likely to be associated with student outcomes. After considering methods such as direct, systematic observations of all teachers, review of local rating reports and tests of cognitive ability, a questionnaire was chosen as the most appropriate, cost-effective method of gathering such data.

For the most part, items were drawn from existing instruments such as those used in the Equality of Educational Opportunity Survey and the ESEA Title I Elementary School Survey, since these items had already undergone a trial and revision process. These instruments were valuable references for information concerning teacher's socioeconomic background, education, professional experiences, and other educationally relevant information.

Original items were written to assess the extent to which teachers updated their skills via inservice training or formal enrollment in institutions of higher learning. Other items were written by project staff to gather data on teacher attitudes toward their current employment status, their preferences for specific types of students or school environment, and their opinions on educational policies.

Comments on the first draft of the questionnaire were solicited from school personnel who would be participating in Project LONGSTEP and from appropriate groups within USOE. Revisions were made on the basis of responses from these sources as well as on reviews by test experts within the AIR organization. As a final step in this early developmental process, the teacher questionnaire was administered to a group of teachers not involved in its development.

Pretest and Revision of the Student Questionnaires and the Teacher Questionnaire

During the 1969-70 school year, the student questionnaires and the teacher questionnaire were pretested in two steps. The first step involved the administration of each questionnaire on an individual basis to a small number of students and teachers in order to elicit their critical appraisal of the questions and to indicate any problems they had in either interpreting the questions or selecting item alternatives.

The second step of the pretest involved a large-scale administration of all forms of the student and the teacher questionnaires in a number of schools in one district that were not part of the main study. The purpose of this endeavor was to obtain response data for item analysis purposes and at the same time check on the validity of comments and criticisms obtained in the first pretest step. The students in this sample represented a cross-section of racial-ethnic and socioeconomic groups.

Although the procedures used varied according to particular classroom situations, in general, an attempt was made to administer the questionnaires under the conditions anticipated for the operational administration of the final version. The major departure from this attempt was that after administering the questionnaires, a period was set aside for discussing or obtaining comments from the respondents. For the younger children, comments and questions were recorded as they arose during administration.

Altogether, a total of 220 students in Grades 3 through 12 in five schools completed the student questionnaire, while background information on an additional 41 students in kindergarten through third grade was provided by six teachers. During the large-scale pretest of the teacher questionnaire, information was obtained from 12 elementary and 40 junior high school teachers. The respondents' comments, the observations of the questionnaire administrators, and the item frequencies and correlations served as the bases for revising both the student questionnaires and the teacher questionnaire. The revised questionnaires were submitted to appropriate federal agencies for clearance.

Educational Treatment Documentation

Introduction. Educational treatments were documented by AIR staff via information gathered from interviews with principals and teachers, from classroom observations, and from existing documentation of the treatment. For purposes of obtaining this information, an educational "treatment" was defined as a unique combination of educational characteristics experienced by one group of students as opposed to another. A treatment group, then, was composed of students who were receiving an educational experience different in some respect from that received by some other group of students. According to this definition, treatments could cut across subject matter areas, grades and even schools. On a practical basis, however, the largest group of students comprising a given treatment unit was the collection of students in one school. An Educational Experience Analysis Guide (EdExAG) was developed to aid AIR personnel in differentiating among various treatment units and in quantifying the groups so defined. The development of this guide and the methodology by which staff documented the educational experiences of students participating in the study are discussed in the following two sections.

Educational Experience Analysis Guide (EdExAG). Undoubtedly, the most challenging task we faced in designing our data collection instruments was preparing a form that could be used to collect information about the wide variety of educational processes occurring in the schools and classrooms in the study. One critical requirement of the study design was that we be able to document the kinds of educational experiences that each student had during the school year. These experiences included characteristics of the curriculum both inside and outside the classroom, and covered such diverse elements as class grouping and structure, extent of individualization of instruction, teaching team composition, scheduling characteristics, community resources used, and amount of teacher inservice training.

It became evident early in the development of this instrument that, although there were a limited number of labels used by school districts to describe their educational activities, in practice there was a wide disparity among educational treatments sharing the same descriptive label. For example, "modular scheduling" is a widely used term for what are in fact various configurations of class schedules. At the same time, programs bearing different labels often turned out to be very much alike. Thus, one district's program of "individualized instruction" could be much the same as another district's "nongraded" program. Therefore, since this study proposed to investigate the relationships between educational practices and various educational outcomes in a higly diverse groups of programs and schools, it was deemed necessary to develop a system by which complex educational experiences could be described and quantified with respect to specific observable characteristics rather than on the basis of variously defined local labels. In other words, what was needed was a conceptual scheme within which educational treatments could be described on the basis of their common denominators. Such a system could serve two major purposes. First, it would provide a framework for the placement of treatment variables in discrete categories so that every student in the study could obtain a score on each variable for use in the statistical analyses planned. Second, it would guide project staff in deciding what information to seek concerning the educational treatment they would be describing, as well as deciding how to record the data collected.

With these major purposes in mind, the staff next formulated some additional desired characteristics of the proposed system. Briefly, these were that:

(1) the system should be as descriptive as possible and utilize a manageable number of items;

(2) the system should not be restricted in its application to just those schools presently in the study, but should be somewhat suitable for general application;

(3) the system, in order not to be biased toward expected effects, should not be based on any single theory of learning or on any particular viewpoint as to the means by which changes in educational activities take place;

(4) each item should be specific enough to permit the detection of qualitative or quantitative differences among educational strategies; and

(5) items and item alternatives should be stated explicitly enough to permit accurate appraisal of educational treatments.

Items were then written so that, when applied to existing school programs, a score could be assigned by the project staff member to document the educational experiences of the children at that school. Alternatives were provided for each item so that it would be possible for AIR staff to locate educational experiences on a continuum ranging from traditional/conventional to innovative or, in some cases, simply to record the presence or absence of certain procedures. Further refinements were made by comparing all items on the basis of their shared characteristics or focus of attention. These comparisons led to a smaller number of items because items that measured to varying degrees some of the same things were combined.

An Educational Experience Analysis Guide (EdExAG) was then created by combining these items to form a single document. Project staff used this initial EdExAG to assist them in their analysis/documentation of the educational activities occurring at participating schools during the 1970-71 school year. Use of this instrument in the field during the first and second years of implementation led to several additions, deletions and refinements of the items. The most notable change was the addition of two tables in which the site visitor was to record (1) the amount of time per day that students in the EdExAG group were taught language arts, math, social studies and science; and (2) the number of years that treatment could have been in effect for the students in the EdExAG group. Items included in the analyses conducted for this report are discussed in greater detail in Chapter III.

Treatment documentation methodology. As noted before, the term "treatment" was used by this study to refer to a unique combination of educational practices and procedures (related to such elements as the curriculum or program objectives, types of materials and how they are used, and class structure and procedures) by which the school environment of one group of students could be differentiated from that of another group of students. Since AIR personnel documenting the treatment in a given school identified groups of students from one school at a time, the data collection unit for treatment data could include, at most, all the participating students in a single school. This data collection unit was termed an Educational Experience Analysis Guide group or an EdExAG group.

The EdExAG groups in each school were initially identified during interviews conducted by staff members with principals and teachers. If necessary, a more detailed interview was conducted with key classroom teachers. At this point, if it appeared that there were significant differences in practices within this group with respect to one or more items on the EdExAG, a second teacher was interviewed. If intragroup differences were confirmed, the initial EdExAG group was split into two groups and a separate EdExAG completed for each. This process continued until it was no longer possible to differentiate among groups of students with respect to the practices and procedures assessed by the items on the EdExAG.

EdExAG groups did not include students from different schools even where groups within two different schools may have had the same scores on all EdExAG items. This permitted the study to be sensitive to possible interschool differences in treatment attributes not documented by the particular items in the EdExAG. Thus, different EdExAG groups represented either unique educational strategies (as defined by the items on the EdExAG) within the same school or treatment approaches in different schools that were not necessarily unique.

Classroom Documentation

The study design included a provision for classroom observation as a means of augmenting and verifying the information about instructional programs obtained through principal and teacher interviews. During the 1970-71 school year a small substudy was conducted to determine the relative effectiveness of the human observer as compared with such modes as fixed and mobile TV cameras. As a result of this pilot study, it was apparent that the human observer would be the most efficient way of documenting classroom processes based on such considerations as cost, ease of implementation in the field, the quality and amount of data that could be collected, and the lack of interference with classroom operations. Plans were made, therefore, that beginning with the 1971-72 school year, AIR site visitors would observe at least one classroom in language arts and mathematics for each EdExAG group.

A number of existing classroom observation instruments and systems were examined, but none of these systems was directly applicable to the purposes of the study. Therefore, a new classroom documentation instrument was developed for field use by AIR staff with these criteria in mind:

- Categories used to describe the classroom should include dynamic teacher/student processes and static or contextual classroom variables.
- Scales should be objective as possible.
- Alternatives in scales should be sufficiently differentiated so that different observers would be apt to assign the same rating in a given situation.
- Items should, for the most part, provide information about the educational treatment not already obtained in other ways.
- The form should be short, easy to read and use, and be amenable to straightforward data reduction.
- Identifying information on the form should link the data to particular educational treatments.

Within this framework an instrument was developed containing 24 items: four of these were designed to verify information obtained by the Educational Experience Analysis Guide; the remaining items provided new information thought to be best collected via observation rather than interview.

Guidelines, definitions and examples were developed and used to train field observers in consistent use of the documentation form. (These guidelines will be deposited in ERIC along with the documentation form at the conclusion of the study.) During training all site visitors participated in four cycles of observation using the draft form and guidelines. The four cycles consisted of videotapes used in the 1970-71 pilot classroom documentation study, and included four different classrooms in two of the cooperating school districts. On-site observations began in the winter of 1971.

Due to cost and scheduling constraints, it was possible to conduct these observations in only one language arts classroom and one mathematics classroom per EdExAG group. Although site visitors selected classrooms in cooperation with the principal of each school, they were specifically instructed to avoid observing in a classroom that they judged might yield biased information. Such would be the case, for example, if the principal had suggested a given classroom because "they are doing something special today." In those situations where the principal gave the site visitor an option to choose one of a number of classrooms, the classroom was selected on a chance basis.

Class Membership Forms and Class Rosters

The study design had specified that the treatment and teacher data associated with each student were to include only those treatments and teachers to which the student was exposed. This was accomplished by obtaining, either from the student directly or from school records, the names of each individual student's language arts, math, social studies and science teachers.

During the first year of full implementation, 1970-71, these data were obtained directly from each student via what was termed a Class Membership Form or Teacher/Class List. (In lower grades these lists were usually completed by the teacher.) During the last two years of data collection, however, only high school and junior high school students filled out these lists. Elementary schools and some junior and senior high schools supplied either class rosters (by subject matter area) or report cards. AIR staff then used these lists to create, for each individual student, a record containing the names of all the student's teachers for each subject.

Class lists or rosters were obtained during both the fall and spring of Years Two and Three. Due to unavoidable delays, these lists were obtained only once during Year One--in January/February 1971, students were asked to list their teachers for the first semester.

Student Questionnaire Verification

Two studies were undertaken to evaluate the dependability of data on the student questionnaire. One was a follow-up survey of the parents of a 2% random sample of students to verify the responses to a selected group of student background items. In the other, follow-up questionnaires were sent to the teachers of a sample of students in Grades 1 through 5 to check on the reliability of responses to items dealing with teacher opinions on selected student characteristics, such as school readiness, social relationships, desire to learn, etc. (Teachers provided this kind of information only for students in Grades 1 through 5.) Both the parent and teacher verification questionnaires were abbreviated versions of those that had previously been filled out by students and teachers, and each was accompanied by a cover and explanation sheet. The degree of correspondence between the student background information given by students (or teachers in the case of first and second grade students) and those given by their parents, and between the answers given by teachers on the two different occasions, was expected to provide some idea of the reliability of the questionnaire replies.

The parent verification questionnaire had two levels corresponding to the two forms of student questionnaires, one for first and second grade students, the other for third through twelfth grade students. The verification questionnaires were field-tested on the parents of nine students from one of the participating sites, who were selected to provide a range of SES and family status. The forms were then reviewed and approved by USOE and the Office of Management and Budget, and the use of an incentive was agreed upon. Each parent who filled out and returned a questionnaire was sent an Eisenhower dollar in a plastic case.

The random sample of students was obtained by selecting 2% of the students in each grade at each participating school, with at least one student selected from each grade within a school. Names, addresses and phone numbers of the parents of these students were obtained, and the questionnaire forms were mailed to parents within a week of receipt of the names. Subsequent mailings to each site followed this schedule: first reminder, two weeks after initial questionnaire mailing; second reminder, two weeks after first reminder; third (phone call) reminder, two weeks after second reminder. Replies were recorded as to date received. For each reminder, a copy of the questionnaire was enclosed, marked so that it could be properly attributed to Wave 1 (initial questionnaire), Wave 2 (first reminder), or Wave 3 (second reminder). For the phone calls, a special format of the questionnaire was designed to allow all the questions to be answered briefly over the phone.

The teacher verification questionnaires for each school were sent to the principal of the school approximately three weeks after initial questionnaire administration. Each questionnaire had the student's name printed on it. The principal was requested to have the questionnaire filled out by the teacher who had filled out the initial questionnaire on that student. The completed teacher verification questionnaires were sealed in plain envelopes and returned to the principal, who then sent all completed questionnaires back to AIR in a large envelope provided for this purpose.

Questionnaires completed during spring 1972 were used in the sampling for the verification study. The results of this substudy will be discussed later.

Data Collection Schedule

In order to collect the necessary data on students, teachers and educational treatments, AIR staff members visited each site three times during each school year. A site visit generally ranged from five to eight days in length. During the visits, AIR staff were involved in a variety of activities, including conducting interviews, making classroom observations, and coordinating the administration of the cognitive tests and student and teacher questionnaires. The schedule of activities accomplished during each of the three visits is shown in Figure II-1.

Data	1970-1971			1971-1972			1972-1973		
	1	2	3	1	2	3	1	2	3
Educational Treatment Documentation									
Interviews	X	(X)	(X)	X	(X)	(X)	X	(X)	(X)
Classroom observation		X			X			X	
Student Questionnaire Administration		X			X			X	
Teacher Questionnaire Administration		X			X			X	
Student Cognitive Test Administration	X^a		X			X			X

1 = first visit, usually late fall
2 = second visit, winter or early spring
3 = third visit, usually late spring

(X) = when necessary

X^a = for the first five sites participating in the study

Figure II-1. Data collection schedule for Project LONGSTEP.

Although initial plans envisioned the administration of tests and questionnaires in the fall and again in the spring each year of the study, the intensity of this schedule imposed a burden on the participating schools and would have added significantly to project costs. Accordingly, the testing and questionnaire administration schedule shown in Figure II-1 was adopted. In addition, although some sites were tested during the fall of the 1970-71 school year, not all participating sites were in the study at that time. Thus the spring 1971 data collection period can be considered the first major data measurement point for complete student and teacher information.

D. Data Management and Quality Control Procedures

Overview of the Magnitude of the Project LONGSTEP Data Base

The data collected for this study are notable with respect to the number of students tested, the number of different data collection forms used, the frequency with which data were gathered, and the manner in which all sources of data--student, teacher and educational treatment--were linked to individual students. Other studies have collected information on more individuals, but for sheer quantity and complexity of data, this study is unique. For example, complete longitudinal information for just one of the 30,000 students in the LONGSTEP data base could involve over 3,500 individual items of data, including:

- Student achievement subtest scores (CTBS raw scores and Expanded Scale Scores) derived from the tests administered during spring 1971, 1972 and 1973 (approximately 30 scores per year or 90 scores total).
- Student questionnaire items collected during each of the study's three years of implementation (approximately 35 items per year or 105 items total).
- Educational Experience Analysis Guide (EdExAG) items for language arts, mathematics, social studies and science subject matter areas. Although these data were gathered yearly, the cross-reference information used to link such data to individual students was obtained once during the 1970-71 school year and once each semester for the 1971-72 and 1972-73 school years. It is possible, then, that approximately 400 EdExAG items could be associated with a given student across all four subject matter areas at any one time, or some 2,000 items total for the five time points.

- Teacher questionnaire items for each of the teachers to whom a given student was exposed. Teacher data were collected yearly, whereas the cross-reference information used to link these data with individual students was collected once during the 1970-71 school year and each semester during the remaining two school years. Since the teacher questionnaire contained approximately 60 individual items and since LONGSTEP's design permitted different teachers for separate subject matter areas, some 240 such items could be associated with each student for each time point, or approximately 1,200 overall. Project LONGSTEP also permitted more than one teacher to be linked to each student for each subject matter area and each time point. Although most participating students had only one teacher per subject matter area, some students had as many as nine. The occurrence of multiple teachers, then, further increased the amount of data in the study's data base.
- Classroom observation data pertaining to each language arts and math EdExAG group. Although gathered only once during each of the 1971-72 and 1972-73 school years, these data were associated with individual students on a semester basis. Therefore, approximately 160 total items of this sort could potentially be associated with each student.

When the multi-item scales (discussed in Chapter III) derived from the individual items on each of these instruments are also considered, the total number of variables potentially available on each participating student (assuming no missing data) increases to approximately 4,000 variables per student.

Project LONGSTEP's Data Management Approach

The amount of data compiled by Project LONGSTEP and its multilevel organization meant that a series of complex operations had to be performed in order to manage (1) the collection of such a large amount of raw data, (2) the translation of that raw data into computer-processable form, and (3) the organization and integration of the various sources of input data into a series of student-level master data files. First, a system for distributing, checking and organizing all data collection forms had to be devised. Second, student records had to be identified in such a way that data across different points in time and across different instruments could be merged. Third, all data had to be rigorously checked for accuracy, corrected where possible and flagged if no corrections were possible. Fourth, the different sources of data had to be organized and placed in master files in such a way that any given variable could be specified at the individual student level. The following sections summarize the procedures that were used to collect, check, organize and integrate the many sources of data gathered by Project LONGSTEP during its three years of data collection.

The logistic support system. The scope of Project LONGSTEP and the large number of schools, grades and students involved over an extended period of time required that a logistic support system be established to maintain control of the ordering, distributing and processing of all data collection instruments. Procedures were established to ensure that adequate supplies (such as tests, related materials and questionnaires) were available well in advance of scheduled data collection periods. These procedures were based on information collected from each district as to the number of schools, teachers and students by grades and classes involved in the data collection. The information was then summarized according to site, and materials either ordered or printed. Adequate extra materials were also included in the requirement estimates to handle such contingencies as changes in enrollment, errors in packaging by the test publisher, and missing information about grades. To assist in these activities, various internal control status sheets were developed for use by AIR staff. In addition, a master logistics status chart, summarizing all relevant information by district, including testing dates, questionnaire administration dates and shipping date requirements, was maintained. In order to ease the burden on local school personnel, all test and questionnaire materials were prepackaged at the AIR facility according to the grade or class breakdown requested by the participating districts.

This was done so that upon arrival the test materials could be distributed easily to the schools and classrooms. Materials were shipped to each site in ample time for distribution to teachers before testing or questionnaire administration. A telephone follow-up provided a check on the arrival of all materials.

In order to faciliate on-site handling of materials by principals, teachers, counselors or other school personnel, supplementary materials were prepared and sent along with the tests and questionnaires. These included:

(1) List of test materials required for each student group to be tested
(2) Checklist for each teacher
(3) Checklist for each school coordinator (usually principal)
(4) List of materials sent to each teacher/school
(5) Practice exercises and instructions
(6) Special coding instructions for answer sheets
(7) Computer-generated, stick-on student ID labels (described in the next section)

During the administration of the tests or questionnaires, an on-site AIR data collection coordinator was available to take care of last-minute changes or shortages. Procedures were devised for use by AIR staff to ensure smooth testing and questionnaire administration and prompt shipment of the materials to AIR.

Longitudinal tracking of participating students. One of the biggest complications in any project which collects student data over several years is how to anticipate the correct school location of participating students from one school year to the next. Rather than wait until the following school year to reorganize data collection efforts, Project LONGSTEP staff went to considerable effort in the spring of each year to document the scheduled movement of participating students from grade to grade and school to school. A detailed chart was prepared showing the participating grade and school combinations for the study. This chart, coupled with the information from the sites about scheduled student movement, for example, from an elementary school to a specific junior high, was used to estimate each student's mobility across grades and schools. In most instances, student movement occurred as anticipated; nearly all students advanced one grade each year, and most stayed within the same school or at least the same district during the three data collection years. The three types of interschool movement that did occur were (1) student movement between participating schools, (2) student movement between participating and nonparticipating schools within the same district, and (3) student movement between participating schools and schools outside the participating district. As specified in the study design, no attempt was made to track individual students unless they moved between participating schools.

Because of the large number of instruments administered during the study, an accurate identification of each student was essential. In order to make this system of identification as complete and as meaningful as possible, a 24-character ID code was created for each student consisting of that student's site code, last name, first name and birthdate. An equivalent 5-digit numeric ID was also assigned to each student. This numeric ID was permanently included with each student's data, since the 24-character ID code was removed from all data files at the end of the study to protect the privacy of each participating student.

Requiring every student to fill out complete identifying information on every data collection form would have placed a needless burden on students and would have been very difficult to implement at the primary grade levels. Moreover, needless inconsistencies (use of nicknames, ignorance of birthdate, etc.) would have resulted from such a procedure. As a means of identifying student-level data with a minimum of confusion and expense, a set of computer-generated, stick-on labels containing all pertinent ID information for each student was prepared and distributed along with the test, questionnaire and teacher identification (i.e., class membership) forms. The labels were placed on the forms at each site.

Instructions relating to the use of identification labels on class membership forms, tests and questionnaires specified that new students or students who did not receive their proper labels were to fill out the identification information on the forms themselves. ID numbers were then assigned to such students in-house before raw data forms were optically scanned or keypunched. Records of who was assigned what number were maintained in order to facilitate subsequent data processing. These records were destroyed at the end of the study.

Processing input data and quality control. After each field administration of a student-level instrument--test, questionnaire or teacher identification (i.e., class membership) form--the raw data forms were shipped back to AIR. An elaborate quality control operation was then initiated to (1) verify that all possible data had been retrieved from the field; (2) check and correct problems in student identification, e.g., new students using the wrong ID code, unanticipated school or grade shifts, etc.; (3) sort all forms alphabetically by name within grade within school; and (4) clean up the forms to facilitate subsequent optical scanning and error correction. Teacher questionnaires were processed in a like manner.

After this initial processing was completed, the raw data forms were converted to machine-readable form by optical scanning or keypunching. The result was a series of intermediate data sets in which each separate record contained (1) an instrument code pointing to the type of data contained on the record and the semester of the study in which it was collected, and (2) an ID code identifying the individual to whom the information on the record belonged. Every machine-processable version of the input data was then computer-scanned to ensure that all values fell within acceptable edit limits. Incorrect records were edited and late arriving forms were added to the files by means of a remote terminal text editing system. Finally, these data were added to the study's master files.

The identification of each student's teachers required a special series of operations. Because of the difficulty in administering class membership forms in Grades 1 through 6, the equivalent information for these grades was obtained through the use of class rosters provided by the participating schools. The names of students and teachers appearing on these class rosters were used by AIR staff to create a data record for each student that looked exactly like a record generated from a class membership form.

Data obtained by AIR site visitors by means of the Educational Experience Analysis Guide (EdExAG) and the Classroom Documentation Form were processed in a manner similar to that for the student and teacher data. Each data collection instrument was checked to see that there were no unanswered items and that the student group to which each form applied was adequately identified. These data were then keypunched and used to create a Treatment Master File by means of the procedures described in later sections. As noted previously, one language arts and one mathematics classroom within each EdExAG group present in the 1971-72 and 1972-73 school years were observed for purposes of completing the Classroom Documentation Form. Therefore, these data were attached to the appropriate EdExAG-group record in the Treatment Master File.

Lastly, although extensive efforts at quality control were made at each stage of processing, there were certain errors in our data that could only be detected when all the data were gathered and complete longitudinal data records were constructed for each student. We were able to find and correct a small number of records where (1) a given student's data had been identified with more than one ID, (2) school and grade inconsistencies occurred among forms, (3) the wrong level of the questionnaire was administered (in which case items identical with the proper level were retained), and (4) the wrong level of the achievement test was administered (in which case the test data were blanked out).

The data from approximately 100 students, however, were so inconsistent that they could not be corrected. Some of these students appeared to be special education cases although not specifically designated as such by their schools. Their data were removed from the study's master files and placed in a separate data file.

In addition, some 170 students changed grades out of sequence. Since these changes appeared to be valid, these students were flagged as such and allowed to remain in the data base.

All procedures were designed so as to maintain the highest level of quality control. Because of such attention to detail, Project LONGSTEP's final data base is a product of several person-years of effort devoted specifically to data cleanup and verification, activities that did not cease until all conceivable sources of misinformation were investigated.

Data file structure. During the design phase of the study, it became evident that a comprehensive and flexible data management system would be the key to effective subsequent analysis of the data. To store, retrieve, and cross-reference extensive information on the nearly 30,000 students, approximately 1,500 teachers, and the many educational treatments occurring in the participating districts necessitated a very sophisticated approach to the organization of the data base. An essential requirement was that each student's performance data, background, attitudes and exposure to educational treatments had to be tied to that student for each year of the study. It was also necessary for teacher and educational treatment data to be independently linked to individual students because classroom membership varied by subject matter area and because the manner in which schools assigned students and teachers or combinations of teachers to different classrooms was complex (e.g., two students may have had different teachers for math but essentially the same treatment, at least as measured by the EdExAG).

In order to maintain the necessary information about students, teachers and educational treatments, several major data files had to be developed as part of the data management system. Three student master files were created:

- a Student Codes Master File which contained the instrument codes of all forms completed by the student, his/her school and grade membership and all other forms of identifying information;
- a Student Test Master File which contained all test data gathered during the study's three years of data collection; and

- a Student Questionnaire Master File which contained all student questionnaire data provided by each student during his/her participation in the study.

Teacher data were maintained in a Teacher Master File, while the treatment data derived from the Educational Experience Analysis Guide and the Classroom Observation Form were placed in a Treatment Master File. Lastly, a series of temporary or intermediate student-level files containing the codes identifying each student's teacher(s) and treatment(s) were generated from the class membership forms/rosters on which each student's teacher(s) were listed. One such file was created for each subject matter area for each time point. These files were later used to link individual students with their teacher(s) and treatment(s). The cross-referencing of teacher and treatment data to individual students led to a number of additional student-level master files in which teacher and treatment data were associated with individual student ID numbers. (The manner in which this cross-referencing was accomplished via these cross-reference codes is described in greater detail in the following section.) Analysis files were then generated by merging all these various student master files and extracting the students and variables of interest in a given series of analyses. Figure II-2 summarizes the organization of Project LONGSTEP's data base.

Cross-referencing procedures. As noted previously, the design of Project LONGSTEP specifically stated that all teacher and educational treatment variables associated with a given student would involve characteristics of only those teachers and those treatments to which that particular student had been exposed. Achieving this degree of precision was complicated by the fact that (1) many students were exposed to more than one teacher or educational treatment, and (2) a group of students who had the same teacher(s) and educational treatment for one subject might not have had the same teacher(s) and educational treatment for other subject matter areas. Such variation in the manner in which individual students were exposed to teacher and educational treatment characteristics meant that it would not be possible to associate treatment and teacher data with students on the basis of grade and school membership information

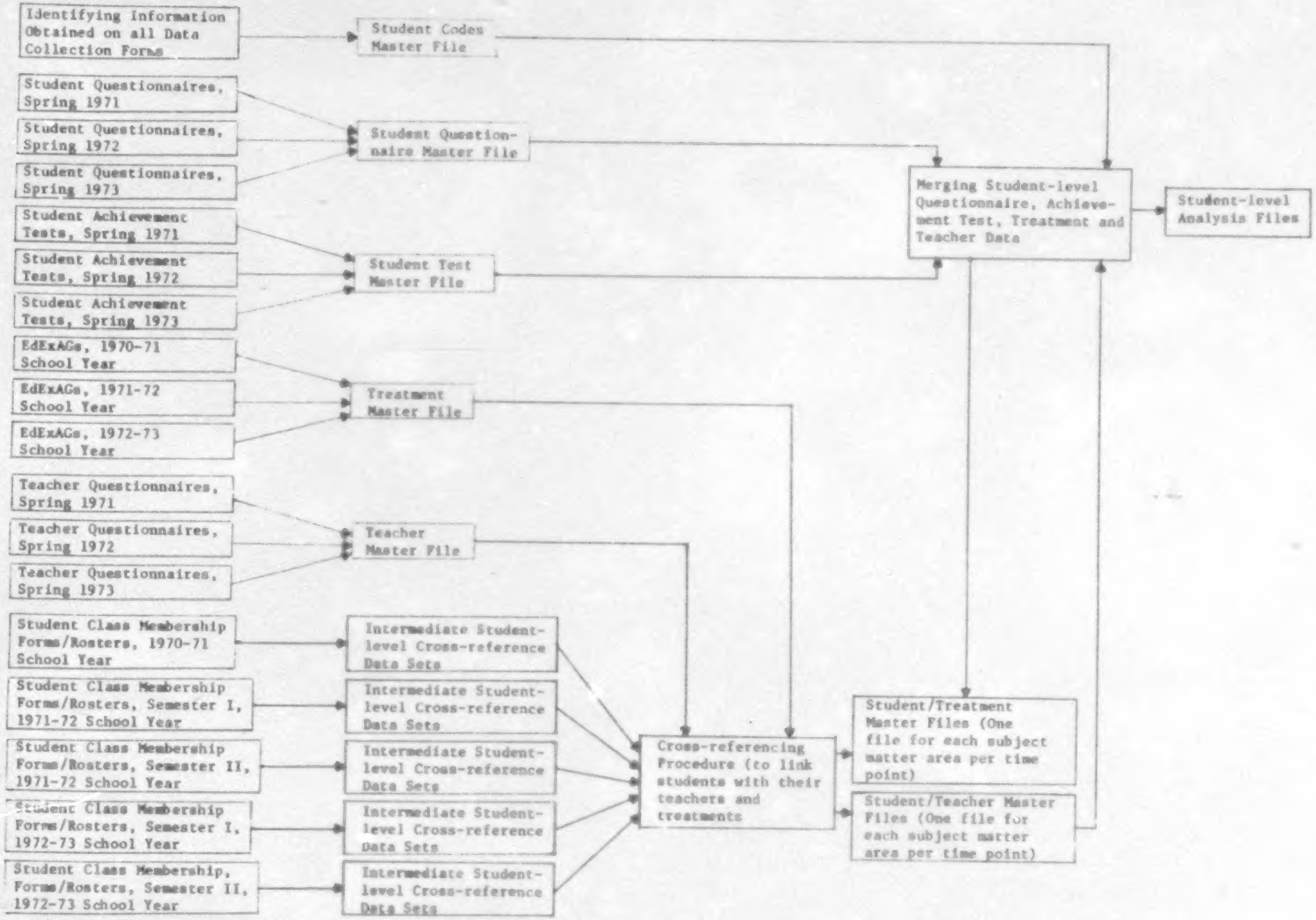

Identifying Information Obtained on all Data Collection Forms
Student Codes Master File
Student Questionnaires, Spring 1971
Student Questionnaires, Spring 1972
Student Questionnaires, Spring 1973
Student Question-naire Master File
Student Achievement Tests, Spring 1971
Student Achievement Tests, Spring 1972
Student Achievement Tests, Spring 1973
Student Test Master File
Merging Student-level Questionnaire, Achievement Test, Treatment and Teacher Data
Student-level Analysis Files
EdExAGs, 1970-71 School Year
EdExAGs, 1971-72 School Year
EdExAGs, 1972-73 School Year
Treatment Master File
Teacher Questionnaires, Spring 1971
Teacher Questionnaires, Spring 1972
Teacher Questionnaires, Spring 1973
Teacher Master File
Student Class Membership Forms/Rosters, 1970-71 School Year
Student Class Membership Forms/Rosters, Semester I, 1971-72 School Year
Student Class Membership Forms/Rosters, Semester II, 1971-72 School Year
Student Class Membership Forms/Rosters, Semester I, 1972-73 School Year
Student Class Membership, Forms/Rosters, Semester II, 1972-73 School Year
Intermediate Student-level Cross-reference Data Sets
Intermediate Student-level Cross-reference Data Sets
Intermediate Student-level Cross-reference Data Sets
Intermediate Student-level Cross-reference Data Sets
Intermediate Student-level Cross-reference Data Sets
Cross-referencing Procedure (to link students with their teachers and treatments
Student/Treatment Master Files (One file for each subject matter area per time point)
Student/Teacher Master Files (One file for each subject matter area per time point)

Figure II-2. General Data management system of Project LONGSTEP.

alone. Information specifically identifying each student's teacher(s) for each subject was gathered to provide this link.

As described earlier, the identification of each student's teachers was obtained in two ways, either directly from a class membership form filled out by students or teachers, or from class rosters supplied by the school. In either case, a data record was generated for each student containing a unique student ID number, school number, grade and teachers for each of the courses in which that student was enrolled (restricted to language arts, math, social studies and science teachers). A 10-digit code was generated from these data to identify each of the student's teachers. A similar 10-digit code was created to identify each student's treatment. This code was of the form AABCCDDEEE, where AA = data-type code, B = subject matter code, CC = student's grade, DD = student's school number, and EEE = teacher number. The data-type code identified the nature of the entity (i.e., the teacher or the treatment) and the school year/semester to which it pertained.

Next, information was gathered from AIR site visitors on the level of specificity needed to identify the treatments received by the students in each grade in each school for each of the four subject matter areas of interest (i.e., language arts, math, social studies and science). When all students in a grade within a school were required to take a given subject and were exposed to the same treatment for that subject, teacher-level specification of EdExAC group membership was not required. In these cases, school and grade information was sufficient to identify the treatment received in that subject matter area. However, for many schools, grades and subjects, EdExAC group membership was associated with specific subsets of teachers. Such was the case, for example, when the students taught by math Team A were exposed to a constellation of educational practices different from those experienced by students taught by math Team B. If teacher-level specification was necessary, then the treatment code for each subject included the teacher number of the student's teacher for that subject. Where the student had more than one teacher for the subject, the lowest teacher number was used to create the cross-reference code. This 10-digit treatment cross-reference code, as already mentioned, was similar in form to that developed to identify a student's teachers.

The teacher and treatment cross-reference codes (in the Intermediate Student-level Cross-reference Data Sets), as discussed in the previous paragraphs, identified the teacher and treatment entities (in the Teacher or Treatment Master Files) to which each student had been exposed. Accurate student-to-teacher cross-referencing was assured by using the same number both to identify teacher questionnaire data and to identify the teachers listed on each student's class membership form. Accurate student-to-treatment cross-referencing was assured by inserting treatment data in the Treatment Master File using IDs that corresponded to the different treatment cross-reference codes found in the Intermediate Student-level Cross-reference Data Sets.

This methodology achieved one of the data management goals of the study by permitting treatment and teacher data to be independently associated with students. The fact that teacher and treatment characteristics were associated with each student so as to reflect the student's own educational environment is novel to large-scale educational studies. It was expected that this rather precise structuring of the LONGSTEP data base would permit student-level analyses that might yield significant new insights into longitudinal educational effects.

III. DISCUSSION OF VARIABLES EXAMINED IN THIS REPORT

A. Introduction

As discussed in Chapter II, Project LONGSTEP was designed to examine interrelationships among three major categories of data: (1) student cognitive and attitudinal outcomes, (2) student characteristics, and (3) educational treatment and teacher characteristics. Obtaining these kinds of information involved the administration or use (by AIR staff) of five types of data collection instruments: student cognitive achievement tests, student questionnaires, a teacher questionnaire, the Educational Experience Analysis Guide, and a classroom observation guide. The procedures by which the analysis variables examined in this report were developed from the items on each instrument are the focus of this chapter. Also included is a brief overview of the analysis data base.

The first section of this chapter contains a review of the standardized achievement test scores selected for analysis. The second section describes the scale analysis procedures used with the student questionnaire and teacher questionnaire data. Also discussed is the methodology used to create the scale scores that were attached to individual students and to individual teachers.

Following these discussions is a summary of the procedures and results of the student questionnaire verification study described in Chapter II.

The next three sections review the development of analysis variables derived from (1) the Educational Experience Analysis Guide, (2) measures of treatment exposure, and (3) a classroom documentation form. Summary statistics for the scales computed from the data gathered each year are presented in these sections.

Then there is a review of the rationale that was used to select the subset of participating students and subject matter areas examined in the analyses conducted for this final report.

The chapter ends with a section that includes (1) a brief review of the methods and rationale by which treatment and teacher data were associated with individual students and (2) an overview of the multilevel organization and the content of the data base examined in the analyses described in later chapters.

It is important to note that the items and scales discussed throughout this report are those involved in the analyses that were conducted specifically for this report. The magnitude of the data collected by this study during its three years of implementation; the complex and multilevel nature of these data; and the costs involved in processing, cross-referencing, aggregating, analyzing and interpreting such a data base dictated the need for a parsimonious approach to this first major analysis of the LONGSTEP data. Therefore, the scale analyses reported here are not exhaustive of all the information obtained on each data collection instrument.

Project staff considered two basic approaches to "streamlining" the data base. One approach was to develop for each instrument composite factor scores accounting for maximum proportions of variance in the items on that instrument. A second approach was to select a limited number of key variables for analysis on the basis of the staff's knowledge of the participating schools, the current research literature, and the analytic goals that the study was designed to achieve. Although the two general approaches are not mutually exclusive, the second alternative was adopted because project staff judged that analyses based on indices computed directly from the items on the data collection instruments would be maximally interpretable, meaningful and useful to the education community. It was also argued that the adoption of a more rational selection of analysis variables would be less likely to capitalize on characteristics of the sites purposely (and not randomly) selected for participation in this study. Lastly, (1) the multilevel nature of the data, (2) the complex manner in which treatment information and data from multiple teachers were independently cross-referenced to individual students, and (3) the use of teacher variables aggregated over each student's possibly unique combination of teachers emphasized the need for using clearly defined and simply derived measurements.

B. Standardized Achievement Test Variables

The schools selected for participation in Project LONGSTEP were chosen because, as a group, they were implementing a variety of different approaches to public school education. Such a diverse sample of schools, therefore, varied greatly with respect to curriculum content, instructional materials and program goals. Consequently, it was necessary for the study to utilize tests of cognitive achievement that not only assessed the kinds of things children learn in school but, at the same time, were relatively independent of the particular content of the materials used for instruction. As discussed in Chapter II, the CTBS (Comprehensive Tests of Basic Skills), Forms Q and R, was the test battery eventually selected. Although it is not curriculum-free, the emphasis in this series is on the "measurement of the group of broad concepts and abstractions developed by all curriculums and on facility in such skills as classifying, manipulating, translating, and interpreting that are required in the effective use of language and number" (CTB/McGraw-Hill, 1969, p. 9).

Since the CTBS was not designed to be used with children in the first grade or in the first half of second grade, an additional test, the California Short-Form Test of Mental Maturity (CTMM-SF), was administered to students at these levels in the 1970-71 school year. Beginning in the 1971-72 school year, however, testing occurred only in spring. Consequently, beginning in the 1971-72 school year, the CTBS was administered at the second grade level and the CTMM was used only at the first grade level.

The test scores that were selected for inclusion in the analyses described in this report are listed in Tables III-1, III-2 and III-3. These tables also show the form and level from which each score was obtained and, for each year of the study, the grades which completed a given instrument.

Table III-4 contains the published reliability coefficients (Kuder-Richardson Formula #20) for each of the subtotal scores of interest for the tests completed for Cohorts 1, 4 and 6. (A cohort is defined by the grade that the student was in or could have been in during Year 1 of the

TABLE III-1

Student Achievement Test Variables, 1970-71

Subtest Name	Test	Form	Level	Grades
MA-Language	CTMM	SF	0	1
			1	2
MA-Nonlanguage	CTMM	SF	0	1
			1	2
Reading Total	CTBS	R	1	3, 4
			2	5, 6
			3	7, 8
			4	9-12
Language Total	CTBS	R	1	3, 4
			2	5, 6
			3	7, 8
			4	9-12
Arithmetic Total	CTBS	R	1	3, 4
			2	5, 6
			3	7, 8
			4	9-12

TABLE III-2

Student Achievement Test Variables, 1971-72

Subtest Name	Test	Form	Level	Grades
MA-Language	CTMM	SF	1	1
MA-Nonlanguage	CTMM	SF	1	1
Reading Total	CTBS	Q	1	2-4
			2	5, 6
			3	7, 8
			4	9-12
Language Total	CTBS	Q	1	2-4
			2	5, 6
			3	7, 8
			4	9-12
Arithmetic Total	CTBS	Q	1	2-4
			2	5, 6
			3	7, 8
			4	9-12

TABLE III-3

Student Achievement Test Variables, 1972-73

Subtest Name	Test	Form	Level	Grades
Reading Total	CTBS	R	1	2-4
			2	5, 6
			3	7, 8
			4	9-12
Language Total	CTBS	R	1	2-4
			2	5, 6
			3	7, 8
			4	9-12
Arithmetic Total	CTBS	R	1	2-4
			2	5, 6
			3	7, 8
			4	9-12

study, 1970-71.) Brief inspection of these coefficients shows that the reliability of each scale is high (.92 or higher for the CTBS scales).

An additional advantage of the CTBS was that the various levels and forms of the test battery had been related to each other and that "a single, equal interval scale across all grades for use with all levels and forms" (CTB/McGraw-Hill, 1970, p. 24) was available. Since students participating in Project LONGSTEP completed alternate test forms each subsequent year of the study and since yearly movement between grades sometimes necessitated the completion of a higher level test, these Expanded Scale Scores seemed particularly suitable for the measurement of longitudinal changes in achievement in basic skills. Development of this scale was based on Thurstone's absolute scaling procedure (Gullicksen, 1950, pp. 284-286) and involved normalizing overlapping distributions of standardization data on the same baseline. The resulting scores were then transformed to a standard score scale with a mean of 600 and a standard deviation of 100 at Grade 10.1 (see CTB/McGraw-Hill, 1970). All analyses conducted for this report and discussed in subsequent chapters involved these Expanded Scale Scores.

So as to estimate the reliability of these Expanded Scale Scores, groups of students in the standardization sample completed more than

TABLE III-4

Published[1] Reliability Coefficients (Kuder-Richardson Formula #20) for Selected Subtest Scores

Cohort	Year	Grade	Test	Form	Level	Reading Total	Language Total	Arithmetic Total
1	1970-71	1	CTMM	S	0		.70[2,3]	.59[2,4]
	1971-72	2	CTBS	Q	1	.95	.94	.92
	1972-73	3	CTBS	R	1	.96	.95	.96
4	1970-71	4	CTBS	R	1	.97	.95	.97
	1971-72	5	CTBS	Q	2	.96	.95	.95
	1972-73	6	CTBS	R	2	.95	.95	.95
6	1970-71	6	CTBS	R	2	.95	.95	.95
	1971-72	7	CTBS	Q	3	.94	.92	.96
	1972-73	8	CTBS	R	3.	.95	.92	.96

[1] CTB/McGraw-Hill. *Comprehensive Tests of Basic Skills - Technical Report*. Monterey, California: CTB/McGraw-Hill, 1970.

[2] From: California Test Bureau. *California Test of Mental Maturity Series, 1963 Revision - Technical Report*. Monterey, California: McGraw-Hill, 1965.

[3] Mental Age - Language

[4] Mental Age - Nonlanguage

one level and/or form of the test. Pearson r's (CTB/McGraw-Hill, 1970) between Expanded Scale Scores based on different levels, different forms, and different levels plus different forms were all high (.80s and .90s). Coefficients as high as these suggest that the CTBS and the Expanded Scale Scores are quite acceptable in terms of reliability.

It was not feasible to compute internal consistency reliability estimates based on Project LONGSTEP data because the costs involved in processing the individual item data so that they would be in a form amenable to analysis would have been prohibitive. It was possible, however, to adjust the published subtest reliability coefficients, based on the assumption that any variance differences between the standardization samples and the LONGSTEP samples were due solely to differences in true score variance (see Gullicksen, 1950, pp. 110-114). The formula for this adjusted coefficient is

$$r_{LL'} = 1 - \left[\frac{s_s^2}{s_L^2}(1-r_{ss'})\right]. \qquad \text{(III-1)}$$

where,

$r_{LL'}$ = the adjusted reliability coefficient;

s_s^2 = the variance of the standardization sample;

s_L^2 = the variance of the LONGSTEP analysis sample; and

$r_{ss'}$ = the reliability coefficient based on the standardization sample.

Table III-5 shows the coefficients that estimate the reliabilities of the Form Q subtest scores (Spring 1972 test administration) based on students in the LONGSTEP sample (in Cohorts 1, 4 and 6). It should be noted that these coefficients are approximations of the desired adjusted reliabilities, that they are presented for comparison purposes only, and that for various technical reasons they should not be used in corrections for attenuation. These limitations occurred for a number of reasons. First, the reliability coefficients reported by CTB/McGraw-Hill (1970) were not based on the entire standardization sample at each grade level. Rather,

TABLE III-5

Adjusted Internal Consistency Reliability Coefficients for the CTBS Form Q

Cohort	Year	Grade	Form	Level	Subtest	CTBS Reliability Estimation Sample N_s	CTBS Reliability Estimation Sample s_s^{2*}	CTBS Reliability Estimation Sample $r_{ss'}$	LONGSTEP Analysis Sample N_L	LONGSTEP Analysis Sample s_L^2	LONGSTEP Analysis Sample $r_{LL'}$
1	1971-72	2.7	Q	1	Reading Total	397	3648	.95	791	4046	.95
					Language Total	392	3844	.94	787	4054	.94
					Arithmetic Total	370	1076	.92	752	1584	.95
4	1971-72	5.7	Q	2	Reading Total	423	6273	.96	1952	4716	.95
					Language Total	421	6691	.95	1995	5817	.94
					Arithmetic Total	422	3697	.95	1943	3190	.94
6	1971-72	7.7	Q	3	Reading Total	429	7621	.94	1520	6044	.92
					Language Total	428	8172	.92	1491	6414	.90
					Arithmetic Total	428	6288	.96	1443	5271	.95

* As noted in the text, these variances were based on the entire standardization sample which had at least 10,000 cases at each grade level.

approximately 400 cases were chosen at each grade level and were used to compute the reliabilities. Second, the variances for these reliability subsamples were not reported in Expanded Scale Score units, the scores on which the LONGSTEP data analyses were based. Third, the only Expanded Scale Score variances that were published in the Technical Report for the entire standardization sample were those for Form Q (CTB/McGraw-Hill, 1970, Tables 9, 10 and 11). However, both Forms Q and R were used during the course of Project LONGSTEP. Although these limitations probably had no substantial effects, it is nevertheless recommended that the adjusted reliabilities reported in Table III-5 be viewed as rough approximations only.

Examination of these adjusted coefficients shows that for each scale, the reliabilities computed on the basis of the LONGSTEP sample were almost as high as those for the CTB standardization sample. (It should be noted that the KR-20 reliabilities reported by CTB/McGraw-Hill [1970] may overestimate the reliabilities of these tests if they are speeded or if the items are not experimentally independent. The published interform reliabilities in scale score units, however, were also in the high .80s and low .90s.)

According to the project's data collection design, the cognitive achievement tests were to be administered as close as possible to the end of the school year at each site. However, practical considerations of scheduling among the participating schools and variations in their starting dates and lengths of school year, meant that a constant time of testing was not feasible. It was therefore necessary to create an analysis variable reflecting such differences so that the test scores could be appropriately adjusted if necessary. This index, Proportion of School Year Prior to Test Administration, is shown for each site for each school year in Table III-6. The impact of these differences will be discussed in Chapter V.

TABLE III-6

Proportion of School Year Prior to Test Administration (By Site)

	1970-71			1971-72			1972-73		
Site*	No. of School Days prior to Testing	No. of School Days per School Year	% of School Year prior to Testing	No. of School Days prior to Testing	No. of School Days per School Year	% of School Year prior to Testing	No. of School Days prior to Testing	No. of School Days per School Year	% of School Year prior to Testing
1	124	180	68.9	120	180	66.7	121	182	66.5
2	131	180	72.8	135	180	75.0	132	180	73.3
3	153	180	85.0	143	180	79.4	141	180	78.3
4	157	180	87.2	147	180	81.7	146	180	81.1
5	157	180	87.2	147	180	81.7	146	180	81.1
6	155	180	86.1	141	180	78.3	141	180	78.3
7	150	178	84.3	125	176	71.0	130	178	73.0
8	159	182	87.4	150	180	83.3	147	182	80.8
9	148	180	82.2	149	180	82.8	146	180	81.1
10	149	178	83.7	135	180	75.0	134	178	75.3
11	153	176	86.9	143	176	81.3	149	176	84.7
12	150	200	75.0	143	200	71.5	169	200	84.5
13	130	180	72.2	138	180	76.7	142	180	78.9

*This site code matches that in the study's computer files.

C. Questionnaire Scale Analysis Procedures and Results

Introduction

The teacher questionnaire and the student questionnaires were designed to measure a number of teacher and student attributes or attitudes which previous research had suggested were educationally significant. A number of these attributes (e.g., sex or ethnic group membership) were categorical in nature and were measured by means of items with nominal response alternatives. However, a number of hypothetical constructs measured by each instrument (e.g., socioeconomic status) were conceptualized as continuous indices possessing interval or ordinal scale properties, and were assessed only indirectly by means of items with nominal or sometimes ordinal response alternatives. The procedures used to develop indices or scales quantifying the constructs of interest in this report and the results of these analyses are discussed next. As noted in Chapter II, copies of the complete instruments will be deposited in ERIC at the conclusion of this study.

Procedures and Results

Many of the items on the teacher questionnaire and student questionnaires had been included so as to provide multiple indirect measures of the same underlying teacher or student attribute. The fact that a number of items were thought to measure a common hypothetical construct or constructs meant that these items should yield higher empirical relationships with each other than with items measuring different constructs. Therefore, the preliminary scale analysis procedures that were implemented on the data obtained from the first full-scale administration of the questionnaires (Spring, 1971) involved the consideration of both rational and empirical interitem associations in the identification of tentative questionnaire scales. Contingency tables were tabulated for all pairs of items. Corresponding contingency coefficients were also computed. A given pair of scores were included in these computations only if both items were nonblank. Project personnel used these interitem contingencies, a priori reasoning, and the scale analysis results reported by Coleman et al. (1966) to identify preliminary single-item and multi-item questionnaire scales and to suggest weights that could be used to transform nominal item alternatives onto an ordinal scale.

Following a review of the tentative scales identified, the items that had not been included in the multi-item scales were reexamined to see (1) if they should be retained as single-item analysis variables, (2) if they were rationally related to any of the multi-item scales of interest, or (3) if they should be dropped from any further analysis. The result of this selective examination of the items contained on each questionnaire was an item pool that measured the student and teacher attributes that project staff believed were the most important for the final report.

Items on each form were then recoded by assigning integer weights to alternatives so that an item would provide a quantitative measure of the construct it supposedly measured. In general, response alternatives were recoded so that higher weights and higher scale scores would indicate more of the construct. The preliminary scale analyses and a priori reasoning, however, suggested that a few items should not be recoded in this manner because such a recoding scheme would lead to nonlinear relationships among scale items. The item on the Form A student questionnaire which asked the teacher to rate the child's social aggressiveness is an example of this kind of problem. Items like this were recoded to give the highest weight to the item alternative reflecting the level of the construct judged as optimal. Thus, average social aggressiveness was assigned a weight of 2, while the alternatives "aggressive" and "passive" were both given a weight of 1. The recodes or item alternative weights that were developed according to these decision rules for the student questionnaires are noted beside the alternatives for each item in Appendix III-A. Teacher questionnaire item recodes are shown in Appendix III-B.

One other decision rule should be noted. On the Form D and Form E student questionnaires two separate items inquired about the occupation of the child's father and mother. If the child indicated that no father was living at home, it was decided that the mother's occupation would be used to indicate the occupation of the head of the household. Project staff believed this decision rule would yield the best estimate of the occupational level of the child's family.

Means and standard deviations were then computed for each recoded item of interest on each questionnaire based on all students, irrespective of grade, who completed a given questionnaire form. These computations were performed separately for the data gathered during each of the three years of data collection. Missing data on each item was defined to be a blank score or a score of zero. Student questionnaire item means and standard deviations are reproduced in Appendix III-C, while the n's are shown in Appendix III-D. Examination of these appendices shows that item means and standard deviations exhibited very little variation across forms and across data collection years. Although the amount of missing data varied a great deal from item to item, for only a few items (e.g., "father's education," Form D, 1970-71) did it go as high as 35%.

Teacher questionnaire item n's, means and standard deviations are in Appendix III-E. These values were also quite stable across data collection years. Teacher questionnaire items contained very little missing data.

Also computed for each year's data were the intercorrelations among these recoded items. Each correlation included only the cases for which there were no missing data on either of the two items being correlated. Such a "pairwise" approach meant that each correlation was based on the maximum amount of information present for a given pair of items. To simplify the examination of these results, the rows and columns of the correlation matrices were rearranged so that the items that formed each tentative scale and the items that might be added to a scale would be in adjacent rows and columns. Each of these correlation matrices was then factor analyzed so that the factorial composition of the items and the scales to be created could be examined. (The fact that none of these matrices yielded negative eigenvalues with a complete principal components solution showed that the effects of the missing item data did not distort the relationships among the items enough to create an internally inconsistent matrix.)

The intercorrelations among the recoded items on each questionnaire and the varimax rotated factor loadings for the matrices computed on each year's data were then inspected by project staff. Final decisions about the item composition of a multi-item scale were made after (1) reaffirming that each scale item was measuring what was judged to be the same construct, and (2) examining the magnitude of the intercorrelations among the items to be placed together in the scale and comparing these intra-scale item correlations with the correlations between these same items and nonscale items. In general, project staff included items in a scale if their highest correlations were with items already in the preliminary scale and if there were no substantial negative correlations with other items in that scale. Each item was allowed to be a member of one and only one scale.

The rotated factor-loading matrices were also examined to see that the items to be included in a scale tended to have similar factorial content. In addition, since scales and time-related changes in them are more interpretable when their items measure a single common attribute, project staff attempted to form scales from items which loaded high on only one factor. This concern with unifactor content was judged to be especially important in the case of the teacher scales because numerous students in the LONGSTEP sample had more than one teacher (for a given subject matter area like language arts or math) and because it therefore would be necessary to compute an aggregate or average teacher scale score for each such student. When teacher scales are of unifactor content, such aggregate scores will tend to be more meaningful and comparable for students exposed to different numbers of teachers who varied with respect to the attribute(s) that the scale was designed to quantify.

The final stage in this scale identification process was to specify those single- and multi-item scales that might possibly be included in project analyses. Table III-7 shows the scale name, number of items, and item content for each of these measures for each student questionnaire form; for the teacher questionnaire scales this same information is shown in Table III-8.

TABLE III-7

Student Questionnaire Scales, Number of Items, Scale Item Composition and Item Numbers

Scale Name	Number of Items	Scale Item	Item Number by Form A^1	B^1	C^2	D^2	E^2
Sex	1	Sex	(1)	–^{3}	(1)	(1)	(1)
Ethnic Group	1	Ethnic Group	(3)	–	(3)	(3)	(3)
Age	1	Age	(2)	–	(2)	(2)	(2)
Socioeconomic Status	3	Head of household's occupation	(6)	(4)	–	(5/6)	(5/6)
		Father's education	(7)	(1)	–	(7)	(7)
		Mother's education	(8)	(2)	–	(8)	(8)
Welfare Assistance	1	Welfare assistance	(5)	(3)	–	–	–
Reading Material Availability	3	Number of books in home	–	–	(15)	(18)	(18)
		Daily newspaper	–	–	(26)	(33)	(33)
		Magazines in home	–	–	(27)	(34)	(34)
Family Structure/Stability	1	Child resides with	(4)	–	(4)	(4)	(4)
Parental Educational Expectations	2	How far in school	–	–	(20)	(23)	(23)
		Parents' desire for excellence	–	–	(25)	(31)	(31)
Attitude toward School	2	Likes school	–	–	(13)	(16)	(16)
		Importance of good grades	–	–	(21)	(24)	(24)
Attitude toward Language Arts	1	Interest in English	–	–	(24)	(30)	(30)
Reading Interest	3	Number of books read	–	–	(7)	(10)	(10)
		Number of library visits	–	–	(16)	(19)	(19)
		Enjoy reading	–	–	(18)	(21)	(21)

TABLE III-7 (continued)

Scale Name	Number of Items	Scale Item	A[1]	B[1]	C[2]	D[2]	E[2]
Attitude toward Math	1	Interest in math	-[3]	-	(23)	(29)	(29)
Importance of School	2	Importance of finishing school	-	-	-	(28)	(28)
		Feelings about stopping school	-	-	-	(35)	(35)
Homework	1	Time spent on homework	-	-	(8)	(11)	(11)
Social Facility	3	Makes friends easily	(9)	(5)	-	-	-
		Social aggressiveness	(10)	(6)	-	-	-
		Confidence with adults	(12)	(8)	-	-	-
Social Identity	4	Liked by classmates	-	-	(12)	(15)	(15)
		Makes friends easily	-	-	(14)	(17)	(17)
		Friends listen	-	-	(17)	(20)	(20)
		Satisfaction with self	-	-	(22)	(26)	(26)

[1] Completed by each student's teacher

[2] Completed by the student

[3] Not asked on that particular form

TABLE III-8

Teacher Questionnaire Scales, Number of Items, Scale Item Composition and Item Numbers

Scale Name	Number of Items	Scale Items	Item No.
Age	1	Age	(17)[1]
Teacher Socioeconomic Background	3	Father's occupation	(4)[1,2]
		Father's education	(5)[1,2]
		Mother's education	(6)[1,2]
Type of Undergraduate Institution	2	Type of institution	(8)[1,2]
		Highest degree offered	(10)[1,2]
Undergraduate Major-English	1	Undergraduate major	(7)[1,2,3]
Undergraduate Major-Mathematics	1	Undergraduate major	(7)[1,2,4]
Teaching Qualifications	5	Highest degree	(15)[1]
		Years teaching experience	(19)[1]
		Years in present school	(20)
		Certification	(21)[1]
		Employment status	(26)[1]
Current Professional Reading	1	Reads journals	(34)[1]
Teacher Associations or Union Participation	1	Professional associations participation	(33)[1]
School Reputation	3	School preference	(27)
		School reputation	(28)
		Opinion of school	(29)

(continued)

TABLE III-8 (continued)

Scale Name	Number of Items	Scale Items	Item No.
Class Size	1	Average class size	(36)
Administrative Support	3	Sufficient school leadership	(32G)
		Class interruptions	(32L)
		Nonteaching tasks	(320)

[1] The item was a non-school specific item and could be used to create a scale score that could be associated with students in other schools in which the teacher might be employed. This occurred only in rare instances where a teacher completed only one teacher questionnaire but taught in more than one school participating in the study during the same school year.

[2] Because this item was not time specific, only one response alternative was associated with each teacher. The alternative that was selected from among the nonblank responses to the item during the three years of data collection was either (1) the most frequently endorsed alternative, or (2) in the case of ties, the most recent response.

[3] Undergraduate Major - English is a dummy-coded (0 and 1) variable that was set equal to 1.0 if a teacher's undergraduate major (item 7) was either elementary education, English or journalism.

[4] Undergraduate Major - Mathematics is a dummy-coded (0 and 1) variable that was set equal to 1.0 if a teacher's undergraduate major (item 7) was either elementary education or mathematics.

Once the analysis variables or scales of interest were determined, the recoded item scores for all students completing a given form of the student questionnaire in a given year were transformed to standard scores, z_i, with mean = 0.0 and standard deviation = 1.0. Items on the teacher questionnaire were also transformed to standard scores for the data from each year considered separately. Students (or teachers) who did not respond to a given item or who had chosen an alternative that was recoded to a blank by project staff (usually the "I don't know" response) were excluded from the standard score computation/transformation for that item.

Scale scores were then computed by summing the standard scores of the nonblank items comprising each scale. Since scale scores were to be formed by summing a varying number of standard scores of the items (because of missing item data) and since this sum (i.e., the scale score) was to have a mean of 100.0 and standard deviation of 10.0, the formula below (after Shaycoft, 1967, p. E-5) was used to accomplish these two computational steps in a single computer run:

$$S_{z_\ell} = 10 \left[\frac{\sum_{i=1}^{n} z_i}{k_n} \right] + 100 \qquad \text{(III-2)}$$

where,

S_{z_ℓ} = the standard score with $\bar{x} = 100$ and $s = 10$ for scale ℓ ;

z_i = the standard score with $\bar{x} = 0$ and $s = 1.0$ for item i ;

n = the number of nonblank items entering the sum;

k_n = the standard deviation of a sum of n standard scores each with $\bar{x} = 0.0$ and $s = 1.0$, or

$$\sqrt{n + 2\left(\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} r_{ij} \right)}$$; or, 1.0 in a single-item scale, and

r_{ij} = the correlation between the i^{th} and j^{th} items that has been based on the observations for which both item i and item j are nonblank.

It is significant to note that this method of computing scale scores from standardized questionnaire item responses was based on means, standard deviations and intercorrelations computed from only nonblank data (or nonblank pairs of scores in the case of correlations). Since each student's or teacher's scale scores were computed from only nonblank item scores, this procedure permitted the computation of scale scores if at least one item from a scale had been answered. For example, if a teacher failed to answer one of the items in a three-item scale, the value of n in Formula III-2 would be 2 and the r_{ij} in the denominator would involve the intercorrelation for the two items the teacher had answered.

The internal consistency of each multi-item standard score scale was then computed to provide an estimate of the reliability of the measurement provided by each. Cronbach's (1951) coefficient alpha was used for this purpose. Alpha is equal to

$$\alpha = \frac{n}{n-1}\left[1 - \frac{\sum_{i=1}^{n} V_i}{V_t}\right] \qquad \text{(III-3)}$$

where,

n = the number of items in the scale;

V_i = the variance of the i^{th} recoded item; or, in this case, 1.0 since all individual items were scaled to $\bar{x} = 0.0$ and $s = 1.0$; and

V_t = the variance of the scale score (i.e., the sum of the items each with $\bar{x} = 0.0$ and $s = 1.0$) prior to standardizing that sum to $\bar{x} = 100$ and $s = 10$.

According to Cronbach, "α estimates, and is a lower bound to, the proportion of test (i.e., scale) variance attributable to common factors among the items" (1951, p. 331). On the other hand, alpha also is said to represent "an upper bound to the concentration in the test (i.e., the scale) of the first factor among the items" (Cronbach, 1951, p.332). Scales with high first factor concentrations will be more interpretable than scales with lower first factor concentrations because it will be easier to see what the items measure in common. The internal consistency measure provided by alpha, however, is appropriate for both unifactor and multifactor scales.

Table III-9 contains the alphas for the multi-item student questionnaire scales. Inspection of this table shows (1) that reliabilities ranged from low (.24) to high (.85), (2) that the only scale whose items have a factual basis (SES) was systematically more reliable than the more attitudinal types of scales, and (3) that the internal consistency of the attitudinal scales seems to be higher for older students. None of these findings, of course, is particularly surprising. One would indeed expect that judgments concerning Parental Educational Expectations, Attitude toward School, Reading Interest and the Importance of School should be more clearly defined for high school students (who completed Form E) than for students of elementary school age (who completed Form C). The higher alphas for the SES scale based on the Form A and B questionnaires were also expected since these forms were completed by each student's teacher, probably on the basis of information maintained by the school or by that teacher. Since there is reason to believe that alpha represents a lower bound estimate of reliability (Lord and Novick, 1968, pp. 87-90), the reliabilities of the student scales were viewed as acceptable, given the particular construct being assessed and the age of the respondents.

Reliabilities for the multi-item teacher scales are shown in Table III-10. These coefficients indicate that three of the scales have moderately high reliabilities (.75 to .82), while two of the scales, Type of Undergraduate Institution and Administrative Support, have low to marginal reliabilities (i.e., in the .40 to .50 range).

TABLE III-9

Coefficient Alphas for All Multi-item Student Questionnaire Scales

		Data Collection Year														
		1970-71 Form					1971-72 Form					1972-73 Form				
Scale Name	Number of Items	A	B	C	D	E	A	B	C	D	E	A	B	C	D	E
Socioeconomic Status	3	.83	.84	-	.77	.72	.84	.83	-	.76	.75	.85	.82	-	.76	.75
Reading Material Availability	3	-	-	.41	.46	.46	-	-	.39	.50	.49	-	-	.41	.47	.51
Parental Educational Expectations	2	-	-	.24	.43	.54	-	-	.27	.44	.52	-	-	.28	.45	.55
Attitude toward School	2	-	-	.40	.48	.58	-	-	.37	.51	.55	-	-	.39	.48	.57
Reading Interest	3	-	-	.41	.64	.67	-	-	.46	.63	.67	-	-	.46	.61	.66
Importance of School	2	-	-	-	.66	.72	-	-	-	.69	.75	-	-	-	.65	.74
Social Facility	3	.55	.56	-	-	-	.53	.58	-	-	-	.52	.57	-	-	-
Social Identity	4	-	-	.37	.37	.37	-	-	.40	.42	.38	-	-	.41	.41	.37

123

124

TABLE III-10

Coefficient Alphas for
All Multi-item Teacher Questionnaire Scales

Scale Name	Number of Items	Data Collection Year 1970-71	1971-72	1972-73
Teacher Socioeconomic Background	3	.79	.80	.79
Type of Undergraduate Institution	2	.49	.45	.38
Teaching Qualifications	5	.82	.81	.79
School Reputation	3	.75	.76	.77
Administrative Support	3	.41	.49	.53

D. Student Questionnaire Verification Study

During the 1971-72 school year, a substudy was undertaken to assess the dependability of the data obtained on the project's student questionnaires. This section contains a brief overview of the methodology and results of this substudy; a more complete description is found in Appendix III-F.

A 2% random sample of students from each participating grade in each school was selected from the population of students who completed a cognitive achievement test in spring 1972. Three hundred forty-two students were selected. Parents of these students were then contacted either by mail or by phone. They were asked some of the same items that had been answered previously by the students themselves, or, for students in Grades 1 through 5, by each student's teacher. Teachers of students in Grades 1 through 5 were also contacted and asked to answer some of the more judgmental types of questions they had answered about each student some weeks before.

Response rates from parents and teachers were very high as more than 93% of the data from each site was obtained. Although only a subset of all of the items on the student questionnaires were examined in this verification study, the results showed evidence of satisfactory dependability of item information whenever the content of the question had a factual basis. On the other hand, there was little parent/student agreement on questions that were interpretive or a matter of personal perception or assessment. A discussion of some of these differences in perception between parent and student and their relationship to student outcomes is presented in Appendix III-F.

Items concerning mother's education, father's education and father's employment showed high agreement between students and parents and between teachers (in the case of the younger children) and parents. Somewhat similar results were also obtained in at least one other large-scale longitudinal study (Echternacht, 1974). The high agreement on these items was particularly noteworthy since they were used to form the study's measure of SES, the major student background index utilized in subsequent data analysis.

E. Educational Experience Scale Analysis Procedures and Results

The Educational Experience Analysis Guide or EdExAG was developed to allow AIR staff to abstract and quantitatively describe a number of significant processes or characteristics of educational programs. This section will briefly review the nature of the data collection unit for this kind of information, the procedures by which the multi-item analysis variables or EdExAG scales were identified and the multi-item and single-item analysis variables developed from this instrument. A copy of the guide used during the final year of project implementation will be deposited in ERIC at the conclusion of the study.

As discussed in Chapter II, the data collection unit for the educational treatment data gathered during the course of Project LONGSTEP was a flexibly defined entity called an Educational Experience Analysis Guide group or EdExAG group. Different EdExAG groups were identified within a school when an AIR site visitor could differentiate among the treatments received by different groups of students by means of items found on the EdExAG. Defined in this manner, EdExAG groups could include as few students as those within one teacher's class for one subject matter area or as many students as those in all grades within a school for all subject matter areas of instruction. EdExAG groups, then, were created in response to treatment variations that existed within a school and within a grade and may be more or less viewed as an efficient but group-level approximation to the documentation of each individual student's educational treatment.

During the 1970-71 school year, this treatment documentation methodology led to the identification of 141 such EdExAG groups, while 167 and 228 groups were isolated during 1971-72 and 1972-73 respectively. These groups, then, represented all the different kinds of educational treatments that were provided by the schools to which students participating in Project LONGSTEP were exposed during each year of the project. Since the EdExAG form was still being developed in 1970-71, the EdExAG data that were obtained during 1971-72 provided the base ingredients for the development of the multi-item EdExAG scales described next.

As with the analysis variables developed from items on the student and teacher questionnaires, the specification of the EdExAG variables to be included in the study's final analyses was based primarily on rational considerations. Senior project staff who were familiar with the sites participating in the study, and with educational programs in general, critically inspected the EdExAG for those items or combinations of items (i.e., tentative scales) that were judged to be the underlying educational characteristics of the various programs included in the study. Item and scale analysis procedures were then undertaken so that the measurement properties of the EdExAG items and the tentative scales might be examined.

The first step in the item/scale analyses of the EdExAG was to rank-order the alternatives for each of the items on an a priori scale from lesser to greater innovativeness, as judged by project staff. Integer weights (1, 2 or 3) were assigned to the alternatives to reflect this rational ordering (see Appendix III-G). Means, standard deviations and frequency distributions for the recoded items were computed, based on the 167 EdExAGs completed during the 1971-72 school year. Recoded items were intercorrelated and the rows and columns of the resulting correlation matrix were rearranged so that adjacent elements of the matrix would contain intercorrelations among items belonging to the same preliminary scale. Senior-level project staff examined these intercorrelations to confirm that scale items were indeed more highly related to other items in the same scale than to items measuring different constructs. Final scale content was determined by considering these empirical relationships and by reaffirming that the items were measuring what were judged to be, on a priori grounds, the same educational construct. Sixteen multi-item scales containing from two to six items were created in this manner. Eleven additional single-item scales were also identified. Table III-11 shows the descriptive label, number of items, and the item content for each of these scales.

Scale scores for these 27 scales were then computed for each EdExAG group by summing the item weights for each of the items in the scale. In order to equate the possible range of scales based on different numbers of items, each scale score was then divided by the number of items in that scale. Since there were no missing item data, no additional adjustments were considered.

Cost constraints did not permit the project to have more than one AIR site visitor documenting each EdExAG group; therefore, it was not possible to compute a measure of the interobserver reliability of the items on the EdExAG. However, two procedures that would tend to produce high interobserver agreement were implemented. First, specific alternatives describing observable school practices were provided for each EdExAG item. Second, prior to site visiting AIR staff discussed each item and developed a relatively consistent interpretation of each. Although these two considerations could not guarantee interobserver reliability, they were procedures that would tend to ensure at least somewhat consistent measurement across observers.

The internal consistency of each multi-item scale was also computed to provide some estimate of the reliability of the measurement provided. Cronbach's (1951) coefficient alpha was used for this purpose. Coefficients for the multi-item scales computed from the EdExAGs completed each school year are shown in Table III-12. These results are presented by year because the EdExAG used during the first year of full implementation did not contain all the items that the later forms contained.

TABLE III-11

Educational Experience Analysis Guide Scales,
Number of Items, Scale Item Composition
and Item Numbers

Scale Name	Number of Items	Scale Items	Item No.
Utilization of Objectives[1]	3	Are there written objectives?	1
		Objectives in behavioral terms	5
		Evaluation used to confirm objectives	71T[2,3]
External Influence on Treatment Characteristics[4]	3	Institution influenced choice of treatment	6[3]
		Consultant's role	8[3]
		Special funding	21[3]
Staff and Materials Evaluation	3	Teacher evaluation procedures	31
		Materials evaluation procedures	32
		Administrative evaluation procedures	73T
Treatment Evaluation	6	Informal classroom tests	66T[3]
		Standardized achievement tests	65S[3]
		Are evaluative procedures systematic?	75T[3]
		Evaluation by administration	76T[3]
		Evaluation by teachers	77T[3]
		Evaluation by researchers	78T

(continued)

TABLE III-11 (continued)

Scale Name	Number of Items	Scale Items	Item No.
Resistance to Treatment[4]	4	Resistance from students	23[3]
		Resistance from teachers	24[3]
		Resistance from administration	25[3]
		Resistance from community	26[3]
School Support Staff	4	Counselors	47
		Learning disability therapists	48
		Home/school liaison staff	49[3]
		Librarian	55[3]
Individualization in Decision Making[5]	3	Topic selection	35
		Materials/text selection	36
		Sequencing decisions	37
Teacher or Locally Developed Materials	2	Externally developed materials	7
		Dependence on locally developed materials	60
Individualization of Instructional Pace[5]	1	Instructional pace	38
Use of Students in Treatment	2	Student helpers	43[3]
		Classroom use of student helpers	44
Utilization of Treatment Evaluation	1	Evaluation used to to modify treatment	72T

(continued)

TABLE III-11(continued)

Scale Name	Number of Items	Scale Items	Item No.
Use of Community Resources	3	Human skill resources	52
		Natural resources	53
		Neighboring institutions	54[3]
Scheduling Characteristics	1	Scheduling	33
Use of External Incentives	1	External incentives	40
Use of Performance Agreements[5]	1	Performance agreements	39
Classroom Group Organization	1	Classroom organization	34
School-Classroom Design	2	Unique architectural features	28
		Classroom structures	63
Teaching Unit Composition	2	Teaching team size	41
		Classroom use of adult aides	42
Status of Treatment Implementation[4]	2	Status of treatment	18[3]
		Longevity of treatment	19[3]
Completeness of Instructional Package[4]	2	Special instructional materials	50[3]
		Use of textbooks	61[3]
Teacher Inservice Training	1	Amount of inservice training	30
Accessibility of Library	1	Accessibility of library	56

(continued)

TABLE III-11 (continued)

Scale Name	Number of Items	Scale Items	Item No.
Use of Media Center	1	Use of media center	57
Utilization of Student Evaluation[5]	3	Classroom use of tests	67S[3]
		Tests used for evaluating student needs	70S
		Evaluation used to modify students' treatment	72S[3]
Affective Evaluation[4]	2	Affective data collected for evaluation (treatment)	74T[3]
		Affective data collected for evaluation (student)	74S[3]
Mechanism for Reinforcement	1	Mechanism for reinforcement[6]	
Opportunity for Practice	1	Opportunity for practice[6]	

---

[1] Primary measures of innovation are underlined.

[2] Items 65 to 78 sometimes applied to the treatment and/or to individual students. A "T" attached to the item number means that the item referred to the treatment while an "S" means that it referred to students.

[3] Item not present on 1970-71 EdExAG and therefore not included in scale.

[4] None of the items present on 1970-71 EdExAG; therefore, the scale could not be computed for 1970-71 EdExAG groups.

[5] Primary measure of individualization.

[6] Found on the attachment to the EdExAG used in 1972-73; therefore, the scale was not measured in 1970-71, 1971-72.

TABLE III-12

Coefficient Alphas for All Multi-item Educational Experience Analysis Guide Scales

Scale Name	Number of Items	Data Collection Year		
		1970-71	1971-72	1972-73
Utilization of Objectives	3	.95[1]	.82	.74
External Influence on Treatment Characteristics	3	NA[2]	.82	.83
Staff and Materials Evaluation	3	.65	.63	.54
Treatment Evaluation	6	NA[1,3]	.73	.76
Resistance to Treatment	4	NA[2]	.71	.56
School Support Staff	4	.64[1]	.59	.43
Individualization in Decision Making	3	.68	.71	.77
Teacher or Locally Developed Materials	2	.69	.68	.74
Use of Students in Treatment	2	NA[1,3]	.80	.79
Use of Community Resources	3	.43[1]	.65	.72
School-Classroom Design	2	.79	.76	.68
Teaching Unit Composition	2	.34[1]	.56	.76
Status of Treatment Implementation	2	NA[2]	.45	.45
Completeness of Instructional Package	2	NA[2]	.69	.63
Utilization of Student Evaluation	3	NA[1,3]	.63	.54
Affective Evaluation	2	NA[2]	.71	.59

[1] Fewer items used in the computation of this scale for 1970-71 than were used for 1971-72, 1972-73.

[2] Items in the scale not asked on 1970-71 EdExAG.

[3] Single-item scale in 1970-71.

It should be noted that the scale construction procedures used for items on the EdExAG were based primarily on staff judgments concerning whether or not items were tapping the same underlying educational construct. Empirical relationships were used to confirm associations expected on a priori grounds, and only substantial contradictions led to modifications of this rationale. Project staff utilized this approach for a number of reasons. First, the schools and grades participating in Project LONGSTEP were purposely selected to reflect a range of types of educational experiences in a range of community settings. The desire that the treatment scales and single-item analysis variables be meaningful for educational programs other than those in the LONGSTEP sample suggested that scale development should not be based solely on associations computed on measurements of the highly select sample of treatments in this study. Second, purely analytic methods such as the computation of orthogonal factor or (rotated) component scores were not used because (1) educational programs are collections of related or correlated component processes and such scores would not reflect the actual confounding and dependencies among such educationally distinct characteristics; and (2) the orthogonality of factor score measures of treatment characteristics would not continue to be present in the student-level data once the EdExAG scores were associated with all the students in each EdExAG group (i.e., once these data were weighted by number of students).

F. Measures of Treatment Exposure

In addition to measures of "qualitative" differences among educational practices and procedures, Project LONGSTEP also collected a number of measures of exposure to instruction or to school staff. These data were collected separately from a "typical" classroom in each grade and subject matter area combination within each EdExAG group. It was possible, therefore, for students in two grades to have been exposed to the same kinds of educational practices (i.e., be in the same EdExAG group) but for differing amounts of time each school day. Table III-13 summarizes these treatment exposure and staff resource indices.

TABLE III-13

Measures of Treatment Exposure and Staff Resources[1]

Treatment Exposure Variables	Staff Exposure Variables
Number of School Days per Year	Number of Student Contact Hours per Week[3,4]
Number of Minutes per Day[2]	Teacher/Student Contact Hour Ratio[3,5]
	Paid Aide/Student Contact Hour Ratio[3,6]
	Volunteer/Student Contact Hour Ratio[3,7]
	Student Helper/Student Contact Hour Ratio[3,8]

[1] These measures were all obtained by AIR site visitors and noted on the Educational Experience Analysis Guide. Time and contact hour data for language arts, mathematics, social studies and science instruction were all obtained separately, even for EdExAG groups incorporating more than one subject matter area.

[2] Asked only on EdExAGs used during 1971-72 and 1972-73.

[3] Asked only on the EdExAG used during 1972-73.

[4] Equal to the number of children in a "typical" classroom multiplied by the number of hours per week for a given subject.

[5] Equal to the number of teachers times the number of hours per week spent in a "typical" classroom, times 100 and divided by number of student contact hours per week.

[6] Equal to the number of paid aides times the number of hours per week spent in a "typical" classroom, times 100 and divided by number of student contact hours per week.

[7] Equal to the number of volunteers times the number of hours per week spent in a "typical" classroom, times 100 and divided by number of student contact hours per week.

[8] Equal to the number of student helpers times the number of hours per week spent in a "typical" classroom, times 100 and divided by number of student contact hours per week.

G. Classroom Observation Scale Analysis
Procedures and Results

As noted in Chapter II, at least one classroom in each language arts and mathematics EdExAG group existing during the 1971-72 and 1972-73 school years was observed by an AIR site visitor. If a given EdExAG group included both language arts and mathematics, separate observations were made of the instruction in each subject. Each observer completed a classroom documentation form and the data obtained were then generalized or attached to all classrooms (i.e., all students) in that EdExAG group. In many instances, site visitors observed in more than one classroom in a given EdExAG group. In these instances, the site visitor was requested to synthesize the observations made in order to produce a single set of observations that could be attached to all students in that EdExAG group.

The development of scales from these data was based primarily on rational considerations. Table III-14 shows the scales and the item content of each scale identified. As with the EdExAG scales, the alternatives for each item were recoded on a three-point scale. Scale scores were computed by summing the (recoded) items in that scale and by dividing the result by the number of items in the scale. This last computational step was implemented in order to make the possible range of scores for all scales equivalent. Appendix III-H shows each scale item and the recoded weights assigned to the alternatives of each item.

Coefficient alpha was then computed for each multi-item scale to provide a measure of internal consistency and reliability. Table III-15 shows these values, all of which were moderate.

TABLE III-14

Classroom Observation Scales, Number of Items, Scale Item Composition and Item Numbers

Scale Name	Number of Items	Scale Items	Item No.
Degree of Grouping	1	Degree of grouping	1
Focus of Activities-Cognitive	1	Cognitive activity	2
Focus of Activities-Affective	1	Affective activity	3
Use of Materials	5	Printed materials	8
		Visual materials	9
		Audio materials	10
		Manipulative materials	11
		Instructional guides	12
Classroom Environment	3	Lighting conditions	13
		Heating conditions	14
		Sound conditions	15
Study Arrangements	3	Space per student	16
		Flexibility of seating	17
		Privacy	18
Access to Resources	3	Audio materials	19
		Visual materials	20
		Supplementary printed materials	21

TABLE III-15

Coefficient Alphas for All Multi-item Classroom Observation Form Scales

Scale Name	Number of Items	Data Collection Year 1971-72[1],[2]	1972-73[2],[3]
Use of Materials	5	.64	.74
Classroom Environment	3	.77	.75
Study Arrangements	3	.73	.73
Access to Resources	3	.73	.75

[1] Based on observations related to 166 language arts or math groups.

[2] The number of observation forms does not equal the number of EdExAG groups for a given year. When an EdExAG group involved both language arts and math, two separate observations were conducted, one in each subject matter area.

[3] Based on observations related to 191 language arts and math groups.

H. Analytic Scope of This Report

Because of time and cost constraints, the analyses conducted for this report and described in subsequent chapters involved the achievement test scores, background variables, attitudes, and teacher and treatment data for only three cohorts of students for the subject matter areas of reading, language and math. (A cohort is a group of students who followed the same grade progression during the three years of the study's implementation. Cohorts are numbered according to the grade the students were in during the 1970-71 school year.) Future reports will cover the analyses for the remaining cohorts as well as the general impact of innovation throughout the school environment (including social studies and science). Cohorts 1, 4 and 6 were selected because (1) they would permit the examination of educational changes occurring during the primary years (for Cohort 1), the elementary years (for Cohort 4), and the years of transition from elementary school to middle or junior high school (for Cohort 6); and (2) the data collection design generally resulted in those cohorts containing proportionately more students and treatments than higher cohorts. Table III-16 reviews the grade-level membership for students in Cohorts 1, 4 and 6 during the three years of Project LONGSTEP data collection.

TABLE III-16

Grade-level Membership for Students in Cohorts 1, 4 and 6

Cohort	Grade-level Membership 1970-71 (Year 1)	 1971-72 (Year 2)	 1972-73 (Year 3)
1	1	2	3
4	4	5	6
6	6	7	8

I. Overview of the Analysis Data Base

Two major factors that made the Project LONGSTEP raw data base so complex were the collection of information at differing levels (on educational treatments, on teachers, on students) and the longitudinal nature of the study which meant that data collection occurred over several points in time. In order to create student-level analysis records, a series of complicated methodological steps were enacted.

First, a set of cross-reference codes was generated so that entities at one level (i.e., teachers and treatments) could be independently linked with entities at another level (i.e., students). Second, all cross-reference codes were checked by computer and the anomalies cross-checked. Third, treatment and teacher scales were created. Fourth, actual cross-referencing was done so that treatment and teacher-level data were associated with individual students. Fifth, aggregate measures (the arithmetic mean) of the teacher scales were computed for each student with multiple teachers. Sixth, all potential analysis data for each student from every source for all time points were placed on a single, large, fixed-length record by merging across student-level data files on the basis of student

ID numbers. If no match occurred, a blank space was left on the record where the data would otherwise have been.

Because of the problems inherent in trying to analyze data gathered at many different grade levels (comparability of test results, differences in questionnaires, variations in educational treatments, etc.), it was decided to conduct separate analyses for each cohort selected for study. The few students who skipped a grade or were held back a grade during the three data collection years were not included in the data sets examined for this report because of the decision to do grade-level analyses.

In addition to the basic information collected on each student, cohort-level data files contained variables created especially for analysis purposes. These included (1) a composite measure (i.e., the arithmetic mean) of each student's SES level computed from the SES scale scores from the three questionnaire administrations, (2) the modal response as to student sex and ethnic group, (3) dummy codes pertaining to treatment group membership, and (4) the percentage of the school year completed prior to the test administration in each site (see Table III-6).

In order to increase the efficiency of specific analysis runs, special data sets were created containing only the students and the variables directly used in those analyses. These data sets took the form of either student-by-student fixed length records or of variable-oriented matrices of cross-products and intercorrelations.

The next chapter discusses the analysis strategies used to relate student outcomes with the other variables discussed in this chapter.

IV. THE ANALYSIS STRATEGY

A. Introduction

As discussed throughout this report, the schools invited to participate in Project LONGSTEP were chosen because, as a group, they represented a range of innovative educational approaches and a range of other educationally relevant characteristics (e.g., socioeconomic level). By design, then, Project LONGSTEP was not an experimental investigation in which AIR could randomly assign students (or student groups) to treatment conditions or in which the nature of the treatments themselves could be controlled by AIR. Rather, Project LONGSTEP was designed to be an associational study in which longitudinal changes in student outcome measures could be related to practices underlying a naturally occurring set of innovative educational approaches.

One aspect of this research approach was that the educational practices in each participating school and the criteria by which students were assigned to treatment groups were entirely under the control of local school personnel. Thus, if anyone were in a position to assume the role of an "Experimenter," it would have been the individual cooperating schools and not AIR. Further, it was more than likely that these schools were relatively less concerned about determining the relationships between components of innovative practices and student outcomes than they were about providing the best education for their students.

During the 1970-71 school year, Project LONGSTEP began collecting treatment, teacher and student data in these schools. Although one selection criterion had been that programs were likely to continue for the projected longevity of the study, it soon became evident that project schools were continually making changes in their educational practices. Some modifications were deliberately planned in advance and were of high visibility, while others were much less obvious. In any event, it was noted that the nature of the treatments present in the LONGSTEP schools appeared to be changing throughout the study's three-year duration.

Educational programs undoubtedly evolve as a function of the educational, social and economic environment of the schools and communities in which they are implemented. Although Project LONGSTEP was not designed primarily to relate longitudinal changes in school practices to the possible reasons for such change, it was felt that an exploratory study of the longitudinal stability of certain kinds of innovative practices would be worthwhile. Since these particular analyses did not relate educational practices to student outcomes, however, their results have not been included in the body of the report but are contained in Appendix IV-A.

The remainder of this chapter is devoted to a discussion of the procedures used to analyze the relationships of student outcomes with the treatment and teacher attributes to which LONGSTEP students were exposed. Also summarized in this chapter are the results of a number of subordinate analyses which examined, for example, the maximum amount of bias due to missing data and the adequacy of certain statistical assumptions or analytic models.

B. Analytic Questions and Strategies

Complex analytical and inferential problems are always present in the analysis of nonexperimental data collected from naturally occurring student groups. Numerous authors have described these issues (e.g., Campbell and Erlebacher, 1970; Campbell and Stanley, 1963; Lord, 1967; O'Connor, 1973a, 1973b; and O'Connor and Klein, 1973), with the general conclusion that accurate estimates of educational treatment effects (i.e., internal validity) might not always be possible because extraneous, uncontrolled and unmeasured factors could "produce effects confounded with the effect of the experimental stimulus" (Campbell and Stanley, 1963, p. 175).

Systematic experiments, however, cannot always be easily implemented in public schools (Evans and Schiller, 1970), particularly since "persons not directly involved in it [i.e., educational evaluation] have developed little appreciation for those events which can invalidate evaluation conclusions" (Lee and Holley, 1975, p. 8). It is also true that the amount of effort required in even attempting to implement systematic, confirmatory experiments may be unwarranted at a particular point in a program of research. In this respect, "quantitative case studies of both archival and purposefully collected correlational data should be uniquely useful in examining important educational phenomena where they occur in nature and in building detailed descriptions of variables attending innovations. Although they will not yield causal generalizations of the experimental sort, they will contribute substantially to the kind of ecologically informed reasoning needed to improve experimentation in the future" (Snow, 1974, p. 288). On the other hand, in light of the difficulties of conducting nationwide experimental studies of school effects, it is simply not realistic to view associational studies as merely a means of improving future large-scale experimentation. Rather, it is more likely that such research will continue to be the major source of information upon which national educational policy is based.

One of the major problems in almost all associational and survey research is that conceptually distinct educational practices cannot be assigned to schools in such a way that their separate effects can be estimated. Rather, the independent variables of interest and other antecedents of achievement will likely be intercorrelated or confounded with each other because the educational programs studied, by design, were not meant to be combinations of unrelated educational practices. The particular pattern of confounding in the data does reflect the organizational and educational structure of the schools being studied and, for this reason, is a result in its own right. However, because confounding can also limit the conclusions regarding the impact of these processes on student achievement, it is usually advisable (1) to assess the degree and nature of the confounding present in the data and consider its possible impact when interpreting other results, and (2) to examine questions of major interest from a number of different analytic perspectives so that conclusions will not be based on a single line of evidence.

Given this general rationale, it was deemed essential that Project LONGSTEP analyses should be as intensive as possible within project time and cost constraints and should involve:

- the specification of a logical sequence of questions which could be asked of the data,
- the careful examination of the results obtained at each stage for possible alternative explanations for the effects observed, and
- the comparison of results obtained by different analytic methods.

The remainder of the chapter discusses these analytic questions and the strategies used to explore them.

The first step in developing the strategies for the analyses presented in this report was to specify the unit of analysis to be utilized in all these procedures. In recent years there has been a great deal of discussion among educational researchers concerning the unit of analysis to be employed in investigations conducted in applied classroom settings of teacher and program effects (Wiley, 1970; Wiley and Bock, 1968). Some researchers have proposed that the classroom unit is most appropriate because (1) it represents the basic organizational and instructional unit of the school and (2) the classroom is usually "the smallest unit of the experimental materials that is ... allowed to react or respond independently during the course of the experiment" (Wiley, 1970, p. 282).

Other researchers have argued that the individual student is the most appropriate unit of analysis since (1) teachers do not teach "classrooms" but individual students, (2) it is an individual student's behavior that is changed as a function of instruction, and (3) student-by-treatment interactions can only be examined when student-level variation is present in the data analyzed.

Project LONGSTEP was designed to relate growth in achievement to attributes of educational treatment, the primary attribute of interest being that of intensive educational innovation. Since treatment data were measured at the Educational Experience Analysis Guide (EdExAG)-group level, it would have been possible to aggregate the student-level data collected (e.g., the test and student questionnaire variables) to the EdExAG-group level and to analyze the resulting aggregated data. Such a procedure, however, would have resulted in a substantial loss of degrees of freedom and would have limited the ability to examine even the most simple multipredictor linear models of growth. Furthermore, since schools are not comprised of random collections of students, the systematic grouping of certain kinds of students into certain neighborhoods and therefore into certain schools would have made it impossible to generalize the findings based on these aggregate data to individual students. The term "ecological fallacy" refers to the bias that can result from making cross-level inferences such as these (Hammond, 1973; Robinson, 1950).

Because of these limitations and because the primary objective of Project LONGSTEP was to determine if intensive educational innovation has a substantial impact on individual student achievement, the unit of analysis used in the analyses discussed in this report was the individual student.

1. What general analytic approach was chosen to assess growth?

Although Project LONGSTEP collected data for three consecutive school years, a number of considerations suggested that the most powerful way to analyze the data for this report was to use two pretest/posttest analyses--Year 1 versus Year 2, and Year 2 versus Year 3. One concern that led to this decision was student attrition; a substantial number of students could not be included if the analyses were to be based only on those students present in both Year 1 and Year 3.

A second concern was that growth between Year 1 and Year 3 could be due to growth either during Year 2 and Year 3, or during both school years. A Year 1/Year 3 analysis could not relate a particular growth pattern in achievement to the characteristics of the educational treatment to which the student was exposed during a given school year. This problem was further complicated by the fact that students were often exposed to different kinds of educational practices and/or to different teachers in two consecutive school years.

One proposed analytic approach was based on the hypothesis that the different Educational Experience Analysis Guide groups isolated by the study actually represented a smaller number of "types" of innovative educational practices. It was argued that if such a typology could be created, it would then be possible to analyze achievement differences among groups of students who had been exposed to different types and combinations of approaches during a one-, two- or even three-year period of time. However, exploratory analyses failed to produce a typology that was stable across school years (Coles, 1975--see Appendix IV-B). This result, as well as other considerations, suggested that the typological approach would probably not prove useful

Another possible analytic approach was to develop a model incorporating two pretest measures and two years of treatment and teacher variables. It might have been possible, for example, to develop a path model relating the Spring 1973 test with SES, the Spring 1971 test, and 1971-72 treatment and teacher data, the Spring 1972 test, and the 1972-73 treatment and teacher data. Because of the complex, multilevel nature of the LONGSTEP data base, however, project staff concluded that such analytic models should not be utilized until the data and the confounding that existed among the various indices were more clearly understood.

Additionally, it was felt that performing two separate posttest analyses had the advantage of making possible the cross-validation of regression weights if the results seemed to warrant such an effort.

Based on all of these considerations, therefore, the Year 1/Year 2 and Year 2/Year 3 analytic approach was chosen as the most desirable.

2. How was the potential analysis sample defined?

Project LONGSTEP involved the collection of questionnaire, test, teacher and treatment data for all children in a specified grade in a given school during each year of the study. While our master files contained students who were present in the sample for any one of the three school years during which data were collected, the project's general analytic approach, as explained under Question 1, was to examine separately the growth in achievement between consecutive years. Consequently, it was decided that the potential analysis sample for a given posttest analysis should contain all those students:

(a) who had followed a consistent grade progression during the school years they were present in the Project LONGSTEP sample (i.e., they advanced one grade each school year), and

(b) who had a nonblank pretest score that was within three grade equivalents of the national norm (published by CTB/McGraw-Hill in the Examiner's Manual for each test).

Application of the first criterion resulted in the elimination of students who skipped a grade or who were held back. Although such students may have been those for whom treatment was particularly effective or ineffective, detection of such a finding was not likely since only 170 of almost 30,000 students did not follow a normal grade progression. The second criterion reflected an earlier decision that test results of students who were more than three grade levels below their peers were not to be included in the study.

Lastly, it should be noted that membership in the potential analysis sample for a given pretest/posttest analysis did not depend on the presence of any data pertaining to the posttest year. Students who had left the sample or who had some form of missing data for the posttest year were identified later during the missing data analyses.

So that a given set of analyses would involve students who were in the same grade and of approximately the same age, the total pool of students was divided into cohorts, that is, groups of students who followed the same grade progression throughout the course of the study. Cohorts were identified by their 1970-71 school year grade level. Cohort 1, for example, included all those students who were in the first grade during Year 1. According to this definition, a student who entered the study for the first time as a second-grader during Year 2 or as a third-grader in Year 3 was also a member of Cohort 1. Analyzing the data by groups of grade-homogeneous students had the additional advantage that all of the students in a given analysis had completed the same level of the test and the same form of the student questionnaire.

As noted at the end of Chapter III, the analyses conducted for this report involved only three cohorts of students: those who were first graders in Year 1 (Cohort 1), those who were fourth graders in Year 1 (Cohort 4) and those who were sixth graders in Year 1 (Cohort 6). Also, the analytic scope of this report was limited to the direct examination of three cognitive dependent variables: the CTBS Reading Total, Language Total and Arithmetic Total Scores. Since separate analyses were to be conducted for each of these dependent variables, the data for each analysis cohort were placed in a separate analysis data set. That is, a given student's Reading Total pretest and posttest, as well as information on his language arts treatment and language arts teacher(s), were placed in one data set; while similar data applying to his achievement, treatment and teacher(s) in mathematics/arithmetic were placed in a separate file. A third file contained each student's Language Total pretest and posttest, language arts treatment, and teacher data.

In summary, the combination of three analysis cohorts (1, 4 and 6), two posttest analyses (Year 1 to Year 2 growth or Spring 1972 posttest analysis; and Year 2 to Year 3 growth or Spring 1973 posttest analysis), and three dependent variables (Reading, Language and Math/Arithmetic) resulted in 18 different data sets that were analyzed.

After each of the potential analysis data sets was created, students who "topped out" on either pretest or posttest were identified and excluded from the analysis sample on the assumption that they represented students for whom the test ceiling was too low. So that a minimum number of students would be excluded, "topping out" was conservatively defined as a perfect score or one incorrect item on a given subtest. The number of students topping out on either pretest or posttest is shown in Table IV-1. It should be noted that the majority of topouts in the Spring 1972 posttest analysis for Cohorts 4 and 6 were students who obtained maximum pretest scores. This was undoubtedly due to the fact that during the 1970-71 school year the fourth grade was the highest grade to complete the CTBS Level 1, and the sixth grade was the highest grade to complete the Level 2. Obviously, this test did not have a high enough ceiling for advanced students in these grades. The Spring 1973 posttest analysis

TABLE IV-1

Numbers of Students in Each Analysis Sample for Cohorts 1, 4 and 6

Cohort	Dependent Variable	Spring 1972 Posttest Analysis				Spring 1973 Posttest Analysis			
		Total[1]	Complete Data[2]	Missing Data[3]	Topouts[4]	Total[1]	Complete Data[2]	Missing Data[3]	Topouts[4]
1	Reading	975	791	184	0	1049	767	282	0
	Language	975	787	188	0	1044	755	288	1
	Arithmetic	953	752	201	0	1012	741	259	12
4	Reading	2535	1952	529	54	2439	1925	499	15
	Language	2546	1995	539	12	2417	1902	493	22
	Arithmetic	2537	1943	527	67	2521	1964	509	48
6	Reading	2183	1520	647	16	1968	1552	405	11
	Language	2138	1491	635	12	1908	1507	401	0
	Arithmetic	2111	1443	614	54	1960	1505	393	62

[1] Number of students with a valid nonblank pretest score and valid grade progression.

[2] Number of students with no missing data with respect to the key analysis variables.

[3] Number of students with missing data with respect to one or more of the key analysis variables.

[4] Number of students with a perfect score or one item incorrect on pretest or posttest.

involved the same level of the CTBS as both a pretest and a posttest measure of achievement. While there may have been a very slight trend for a majority of the maximum scores to have occurred on the posttest, students who topped out on either pretest or posttest tended to be those students who obtained near maximum test scores both years. Examination of Table IV-1 shows that the number of students topping out was relatively small, especially when compared to the number of students in the analysis samples. It is not likely, therefore, that elimination of students with perfect or near perfect test scores could have had a large impact on later findings. The other entries in Table IV-1 refer to missing data and will be discussed in the next section.

A check was also run to ensure that no student "bottomed out" on the posttest, i.e., had a score of zero or one correct item on a subtest. In fact, no such students were found.

Other criteria were also used to select the students for the analysis sample examined for this report. These criteria will be discussed in the following section.

3. How were missing data handled and how did students with missing data differ from those with no missing data?

Although a number of different methods to replace or adjust for missing data have been proposed over the past few years, most of these procedures are based on the assumption that the occurrence of missing data is random or that students with missing data do not differ systematically from those on whom the missing data analyses are based (at least with respect to the processes being studied). These assumptions are not plausible in applied educational/social settings and, from a practical standpoint, are largely untestable. It was decided, therefore, that students with missing data on key variables would be deleted from a given analysis sample and that the maximum possible bias to the mean and the variance that could result from this procedure (assuming normality of the complete distribution) would be estimated. An advantage of including only those students with complete data in the analysis samples was that all correlations among the variables would then be based on the same students. Since matrices of such correlations are, of necessity, internally consistent (or Gramian), the problems which occur in the analysis of inconsistent correlation matrices (e.g., the possibility of a multiple R greater than 1.00) could therefore be avoided.

Because of time and cost constraints, missing data analyses were implemented on only three of the eighteen analysis samples discussed in this report. This seemed entirely adequate, however, since there was no reason to believe that missing data were differentially associated with the data collection year or with the different CTBS subtests.

Table IV-1 shows the number of students in each analysis data set who had *any* form of missing data with respect to the key variables examined for this report (see Question 6 for a list). Examination of this table shows that the number of students deleted from the analysis samples ranged from approximately one-fifth to one-third of the available students. Although the numbers of students lost appear to be fairly large, it must be emphasized that, in addition to normal and expected attrition between school years, these figures also reflect any and all forms of missing data in an analysis record that involved a number of different sources of information on each student. Thus, a given student could have been deleted from a given analysis data set because the school had not identified that student's teacher(s); the student's teacher had not completed a teacher questionnaire; the student had been ill during the administration of the appropriate subtests; the student had left the school; or he/she was enrolled in a non-grade-level treatment (e.g., a fifth grader in a fourth grade class), etc.

These various sources of missing data were not delineated because the goal of the missing data analyses was not to speculate on the possible reasons for loss of data but rather to determine the overall consequences of deleting a subset of students from the analysis. Therefore, the missing data analyses that were conducted involved comparing the students deleted because of missing data with the students included in the analysis sample.

Estimates of bias were calculated according to formulas given by Kelley (1947, pp. 296-298) that are based on the assumption that the attribute in question is normally distributed. Bias to the mean (if the missing data constitute a tail of a normal distribution) is $d_m = ky/p$, where $k = 1$ if the observations are lost from the negative tail and -1 if the observations are lost from the positive tail; y is the ordinate of the normal distribution at the point of truncation and p is the proportion of the normal distribution remaining after truncation. It should be noted that unless all the students with some form of missing data constitute a tail of the posttest distribution, d_m (as defined here) provides an upper limit to the bias rather than an estimate of it.

The variance of the truncated distribution is $V = 1 + (yz/p) - d_m^2$, where y, p and d_m are defined as before and z is the normal deviate associated with p.

The three analysis data sets that were selected for the missing data analyses were the Spring 1973 Reading Total Score analysis samples for Cohorts 1, 4 and 6. Examination of Table IV-1 shows that the number of cases lost because of some form of missing data during the 1973 posttest analysis amounted to 27%, 21% and 21%, respectively, of the total possible number of cases present in each of these data sets. (Percentages were computed by dividing the number of cases lost due to missing data by the total number of students, excluding those who had topped out.) Table IV-2 shows the mean and variance of a truncated distribution of scores after 27% or 21% of the cases have been deleted from the negative tail of an originally normal distribution (using the formulas given by Kelley, 1947). The mean and variance of the complete normal distribution prior to truncation are $\mu = 0$ and $\sigma^2 = 1.0$.

The last two columns in Table IV-2 show the maximum bias to the mean and variance if the specified proportion of cases deleted constituted the negative tail of a normal distribution. Bias was computed by subtracting the known value ($\mu = 0$, $\sigma^2 = 1.0$) from the value for the truncated distribution. This table suggests that the maximum amount of bias to the Reading Total posttest score mean could be almost one-half of a standard deviation (assuming $\sigma^2 = 1.0$) for Cohort 1; and, that a

TABLE IV-2

Means (μ') and Variance (σ'^2)
of a Truncated Normal Distribution ($\mu = 0$, $\sigma^2 = 1.0$)

Cohort	% of Cases Deleted	% of Cases Remaining	Mean and Variance of the Truncated Distribution		Bias	
			μ'	σ'^2	to μ	to σ^2
1	27	73	.45	.52	.45	-.48
4,6	21	79	.36	.58	.36	-.42

48% reduction, at most, in the total variance could have resulted in the Cohort 1 analysis.

The maximum amount of possible bias to the mean and variance as shown in Table IV-2 is fairly moderate. It is extremely unlikely, however, that all of the students deleted because of missing data actually represented extreme scorers.

It should be noted that these analyses did not directly address the issue of the possible impact of missing data, whatever the cause, on treatment inferences. Although it may seem like a fairly straightforward question, analyzing the potential impact of missing data is extremely complicated and, as mentioned previously, highly dependent on assumptions that are not particularly reasonable with respect to the social and educational processes involved in public school education. Even the simple question, "how has the correlation between a treatment variable and posttest been affected by the loss of students with missing data," is not amenable to easy examination; nor would such examination yield meaningful results. This is because all the analyses that could be conducted would be based on students who have nonmissing data, and there would be no reason to believe that these students were representative of the students with missing data. For example, it would be possible to compare the treatment/posttest correlation based on the students in an analysis sample (i.e., those with no missing data) with the correlation based on all those students who had some form of missing data but complete data with respect to the treatment and posttest. It is not likely, however, that the results of such an analysis could be generalized to include those students who left the schools and the LONGSTEP sample between the pretest and posttest school years and who represented the majority of cases in the missing data sample.

The impact of attrition also could not be studied directly because both the posttest and the posttest-year treatment score were missing. In this case, it would have been necessary to compare the correlation of pretest and pretest-year treatment for students with and without data for the posttest year. However, generalizing the results obtained from an analysis of pretest-year data to the possible impact of attrition on the posttest/treatment correlation would not be particularly valid since (a) students were not necessarily exposed to the same kinds of treatments and teachers during the pretest and posttest school years and (b) the students involved in the missing data analyses would have been a year younger and at an earlier point in their educational careers.

The situation described here would become impossibly complex if the same question were to be asked of the possible influence of missing data on a multipredictor linear model of growth in achievement. Since the issues involved in estimating the potential impact of missing data were so complex that even the simplest question would have required considerable intensive investigation and would have led to results of questionable validity, project staff determined that it would not be cost effective to conduct such analyses for this report. Rather, it was decided that a realistic approach would be (a) to examine the differences between the analysis and missing data samples with respect to a selected set of achievement, attitudinal and demographic variables and (b) to restrict our inferences to the kinds of students present in the analysis samples.

Tables IV-3, IV-4 and IV-5 contain the means and standard deviations (unbiased) on a selected set of variables for the analysis sample, for the missing data sample, and for all cases combined with respect to the Spring 1973 (posttest) Reading Total Score analysis data sets for Cohorts 1, 4 and 6. These tables contain both demographic and selected

TABLE IV-3

Means, Standard Deviations, and Point-Biserial Correlations (r_{pb}) for the Missing Data Analysis of Cohort 1, Spring 1973 Reading Total Score Analysis Sample

Variable[1]		Analysis Sample (N=767)	Missing Data Sample (N=282)	Combined Sample (N=1049)	r_{pb}
Reading Total Score[2]	$\overline{X}$	338.46	329.55	336.06	
	SD	63.97	63.46	63.93	$.06^{*}$
	N	767	282	1049	
Language Total Score[2]	$\overline{X}$	333.15	319.84	329.58	
	SD	65.28	62.78	64.86	$.09^{**}$
	N	761	279	1040	
Arithmetic Total Score[2]	$\overline{X}$	319.53	313.20	317.85	
	SD	39.26	40.10	39.56	$.07^{*}$
	N	733	265	998	
Level of Innovation	$\overline{X}$	20.13	19.32	19.91	
	SD	4.16	4.46	4.26	$.08^{**}$
	N	762	240	1037	
SES	$\overline{X}$	100.58	96.61	99.64	
	SD	9.98	9.55	10.02	$.17^{**}$
	N	767	240	1007	

[1] Data used for these analyses were obtained during the pretest year.

[2] Not adjusted for time of testing differences among sites

*$p < .05$ (two-tailed)

**$p < .01$ (two-tailed)

TABLE IV-4

Means, Standard Deviations, and Point-Biserial Correlations (r_{pb}) for the Missing Data Analysis of Cohort 4, Spring 1973 Reading Total Score Analysis Sample

Variable[1]		Analysis Sample (N=1925)	Missing Data Sample (N=499)	Combined Sample (N=2424)	r_{pb}
Reading Total Score[2]	$\bar{X}$	475.73	465.57	473.63	
	SD	66.96	70.41	67.80	.06**
	N	1925	499	2424	
Language Total Score[2]	$\bar{X}$	479.59	463.24	476.27	
	SD	75.20	73.60	75.15	.09**
	N	1925	491	2416	
Arithmetic Total Score[2]	$\bar{X}$	452.01	441.63	449.89	
	SD	56.66	59.41	57.37	.07**
	N	1914	492	2406	
Attitude toward Language Arts	$\bar{X}$	99.71	99.47	99.67	
	SD	9.63	9.58	9.62	.01
	N	1706	358	2064	
Reading Interest	$\bar{X}$	99.01	100.18	99.21	
	SD	10.36	10.18	10.34	-.04*
	N	1708	358	2066	
Attitude toward Math	$\bar{X}$	99.41	100.27	99.56	
	SD	9.66	10.26	9.77	-.03
	N	1703	355	2058	

(continued)

TABLE IV-4 (continued)

Variable		Analysis Sample (N=1925)	Missing Data Sample (N=499)	Combined Sample (N=2424)	r_{pb}
Social Identity	$\overline{X}$	101.37	100.77	101.26	
	SD	9.86	9.68	9.83	.02
	N	1708	358	2066	
Level of Innovation	$\overline{X}$	18.74	18.29	18.65	
	SD	4.29	4.67	4.37	.04*
	N	1924	495	2419	
SES	$\overline{X}$	99.72	99.39	99.66	
	SD	10.03	10.16	10.05	.01
	N	1925	444	2369	

[1] Data used for these analyses were obtained during the pretest year.

[2] Not adjusted for time of testing differences among sites

*p <.05 (two-tailed)

**p <.01 (two-tailed)

TABLE IV-5

Means, Standard Deviations, and Point-Biserial Correlations (r_{pb}) for the Missing Data Analysis of Cohort 6, Spring 1973 Reading Total Score Analysis Sample

Variable[1]		Analysis Sample (N=1552)	Missing Data Sample (N=405)	Combined Sample (N=1957)	r_{pb}
Reading Total Score[2]	$\overline{X}$	545.45	527.39	541.71	
	SD	72.45	68.22	71.95	.10**
	N	1552	405	1957	
Language Total Score[2]	$\overline{X}$	537.11	518.20	533.21	
	SD	79.33	78.20	79.45	.10**
	N	1545	401	1946	
Arithmetic Total Score[2]	$\overline{X}$	522.33	499.62	517.72	
	SD	75.23	71.33	75.00	.12**
	N	1532	390	1922	
Attitude toward Language Arts	$\overline{X}$	100.37	100.07	100.32	
	SD	9.99	10.56	10.09	.01
	N	1349	272	1621	
Reading Interest	$\overline{X}$	100.31	100.22	100.29	
	SD	10.05	10.63	10.14	.01
	N	1351	273	1624	
Attitude toward Math	$\overline{X}$	100.17	98.55	99.89	
	SD	9.92	9.59	9.88	.06*
	N	1345	272	1617	

(continued)

TABLE IV-5 (continued)

Variable		Analysis Sample (N=1552)	Missing Data Sample (N=405)	Combined Sample (N=1957)	r_{pb}
Social Identity	$\overline{X}$	100.77	99.40	100.54	
	SD	9.64	10.40	9.78	.05*
	N	1350	273	1623	
Level of Innovation	$\overline{X}$	19.47	18.02	19.18	
	SD	4.67	4.73	4.72	.12**
	N	1527	383	1910	
SES	$\overline{X}$	100.06	100.98	100.21	
	SD	9.89	9.70	9.86	-.04
	N	1552	301	1853	

[1] Data used for these analyses were obtained during the pretest year.

[2] Not adjusted for time of testing differences among sites

*$p < .05$ (two-tailed)

**$p < .01$ (two-tailed)

attitudinal variables for Cohorts 4 and 6. Only the demographic information was available for Cohort 1 (i.e., Table IV-3) since items assessing attitudes were not asked of first and second grade students. The measures summarized in the tables were collected during the pretest year. These particular data, rather than those obtained during the posttest school year, were summarized because, as noted previously, posttest-year information was not available for students who had transferred out of the school and therefore out of the LONGSTEP sample. Although a valid pretest (i.e., Reading Total Score) was present for all students in both the analysis and the missing data samples, students in either sample could have had missing data with respect to other measures gathered during the pretest year (excluding SES, which was required to be nonblank for all students in the analysis sample). For this reason, the number of students upon which each mean and standard deviation was based is also shown.

Tables IV-3, IV-4 and IV-5 also contain the point-biserial correlation between each variable of interest and a dummy variable (X) encoding presence (X = 1) or absence (X = 0) in the analysis sample. Since the point-biserial correlation is a function of both mean differences and the relative number of cases in the two samples, it provided a more appropriate measure of the magnitude of the differences that actually existed between the students in these two groups. These point-biserial correlations can be interpreted as estimates of the actual bias to the posttest mean attributable to missing data if it is assumed that the difference between the analysis sample and the missing data sample pretest means adequately represents the difference that would have occurred on the posttest had it been possible to collect such data from these students. (Although missing data due to normal attrition were not identified separately, inspection of student records showed this to be the major underlying reason for missing posttest-year data.)

Analysis sample/missing data sample differences with respect to the pretest score measures were statistically significant but relatively small for all three cohorts. In all instances, missing data sample pretest score means were lower than those for the analysis sample. It was interesting to note, however, that there was a slight trend toward larger achievement score differences for older students. In addition to these differences in achievement, it was important to determine if students deleted from the analyses differed greatly from the analysis sample students in other respects. Although many of the differences reached statistical significance because the total N's were large, Tables IV-3, IV-4 and IV-5 show that there were no large differences between the analysis and missing data samples with respect to most of the other variables. There was a slight tendency for students in the analysis samples to have more positive attitudes and to be of higher SES than the students with missing data, although only those differences for SES for Cohort 1 were notable. Students deleted because of missing data during the posttest year also tended to be exposed to somewhat less innovative treatments during the pretest year than did students in the analysis sample. This result probably reflected the trend that Level of Innovation (see question 6) is, in general, positively correlated with SES and that lower SES families tend to be relatively more transient than higher SES families.

Table IV-6 shows that presence of missing data was not highly related to ethnic group membership (white or nonwhite) or sex. There was, however, a very slight trend for the analysis samples to have a larger proportion of females and whites than the missing data sample.

All these results suggest that the students who were present in the analysis samples did differ from those who were deleted because of missing data. Although the maximum bias to the mean and variance that would result from deleting the numbers of students shown in Table IV-1 could have been moderately large (see Table IV-2), an examination of the analysis sample and the missing data sample in terms of mean differences on the pretest (Tables IV-3, IV-4 and IV-5) suggests that the actual bias was undoubtedly much smaller. As discussed previously, since there really is no way of determining how students who left a school prior to the posttest year (the major reason for missing data) would have done had they remained, it seems safest to restrict our inferences only to the kinds of students who were in the analysis sample (see Tables IV-3, IV-4 and IV-5). On the other hand, given the assumption that presence or absence of

TABLE IV-6

Sex and Ethnic Group Membership for the Cohort 1, 4 and 6 Analysis Samples and Missing Data Samples

		Sex				Ethnic Group			
		Male		Female		White		Nonwhite	
Cohort	Sample	N	%	N	%	N	%	N	%
1	Analysis	360	47%	401	53%	643	84%	119	16%
	Missing Data	151	56%	119	44%	220	81%	50	19%
4	Analysis	948	50%	952	50%	1726	91%	171	9%
	Missing Data	228	53%	206	47%	381	88%	53	12%
6	Analysis	731	47%	821	53%	1360	88%	184	12%
	Missing Data	169	50%	166	50%	280	84%	53	16%

NOTE: Since nonblank data with respect to sex and ethnic group membership were not criteria for inclusion in an analysis sample, the N's in this table may not add to the totals shown in Table IV-1.

complete data does not substantially moderate the relationship between educational outcomes and their correlates, inferences based on the analysis of students with complete data could, in fact, apply to all students. The observation that the students in the analysis samples and the missing data samples did not exhibit dramatic differences on a majority of the variables examined here implies that this may be a reasonable assumption.

4. What was the overall impact of differences in time of testing and what adjustments were made to pretest or posttest scores?

Because of scheduling problems, differences in starting dates and number of school days per year, as well as other considerations, it was not always possible to administer the achievement tests at the same relative point in the school year at each site. Therefore, each pretest and posttest was regressed on the proportion of school year which had elapsed prior to testing. The resulting regression constants were then used to residualize each student's pretest score and posttest score according to the following formula:

$$Y'_i = (Y_i - \hat{Y}_i) + \overline{Y} \qquad \text{(IV-1)}$$

where,

Y'_i = pretest or posttest score for student i that has been residualized for time of testing;

Y_i = original pretest or posttest score for student i;

$\hat{Y}_i$ = predicted pretest or posttest score for student i based on the regression of a score on its site-level time of testing measure; and

$\overline{Y}$ = the arithmetic mean of the original pretest or posttest scores.

The last term in Equation IV-1 was added so that the mean of the residual scores would be equal to the mean of the original scores.

The pretest and posttest scores in each analysis data set were residualized separately. These "adjusted" test scores were used in all subsequent analyses. Table III-6 in the previous chapter shows that the timing of test administrations differed by as much as 30 school days. Nevertheless, in all cohorts and for all analysis data sets the correlations between test scores and time of testing were small. (Nine of the

36 correlations were between .15 and .22, with the remaining 27 ranging from -.07 to .10.) Appendix IV-C shows the means, standard deviations and intercorrelations among the test variables, the time of testing measures, and SES for each analysis data set.

5. What was the functional relationship between pretest and posttest?

Since the models to be applied at later stages in the analysis would either focus on deviations from expected posttest performance or would include pretest as a measure of initial status in a linear growth model, it was deemed advisable that the functional relationship between pretest and posttest be examined prior to conducting these more complex analyses. The primary reason for this concern was that a linear pretest/posttest function would systematically underpredict or overpredict posttest performance for low or high pretest students if the relationship were really curvilinear (Pohlman and Newman, 1973). The nature and extent of the problem would, of course, depend on the particular functional relationship present in the data.

Because the presence of curvilinear components would not be of interest unless they were substantial enough to be readily apparent on visual inspection, curvilinearity was not tested statistically. Rather, deviations (i.e., residuals) from the pretest/posttest linear regression line were plotted on predicted posttest and the resulting scatterplot examined for departures from linearity. Examination of this plot also allowed project staff to determine if variation of the residuals about the regression line was relatively constant throughout the range of pretest or, in other words, whether or not the distribution of residuals was homoscedastic.

Appendix IV-D contains these scatterplots for each of the analysis data sets. Inspection of these plots shows that curvilinear components were not readily apparent in any of the pretest/posttest regressions. (The grouping of scores into arrays on some of the plots reflected the use of CTBS Expanded Scale Scores.) For this reason, it was concluded that this relationship between pretest and posttest was linear in each of the analysis data sets.

6. Which key variables were examined?

Project LONGSTEP was designed to relate four basic categories of data--home background, initial achievement status, educational treatment, and teacher characteristics--to posttest performance. Although it would have been possible to develop an analytic approach containing a large number of variables from each of these categories, project staff felt that the most meaningful and interpretable results would be obtained if these initial analyses were limited to certain key variables. Table IV-7 shows the indicators that were selected.

Items comprising the student SES scale had been asked on the questionnaires administered each spring. Hence, it was possible to average all available SES scales and form a composite measure of SES. It was this aggregate measure of socioeconomic status that was used in all subsequent analyses. The first ten educational treatment indicators shown in Table IV-7 were those Educational Experience Analysis Guide (EdExAG) scales that project staff had previously judged to be primary indicators of educational innovation. The EdExAG scales "Degree of Individualization" and "Level of Innovation" were two superindices formed from the first ten EdExAG scales shown in the table. Degree of Individualization was formed by summing together the four EdExAG scales which were judged to be components or forms of individualized instruction--Individualization in Decision Making, Individualization of Instructional Pace, Use of Performance Agreements, and Utilization of Student Evaluation. Level of Innovation was merely the sum of the first ten EdExAG scales. The last educational treatment indicator was a measure of exposure to treatment and provided a quantity-of-schooling dimension. All these measures were based on the EdExAGs completed during the posttest year for each student's language arts or mathematics treatment (depending on the particular dependent variable).

The teacher characteristic indicators included in Table IV-7 were the mean teacher scale scores for all of a student's language arts or math teachers (depending on the dependent variable being analyzed). The use of such aggregate scores was necessary because some students were exposed to more than one teacher per subject matter area. (The fact that

TABLE IV-7

Key Predictor Variables

Category	No. of Variables	Predictor Variable
Home Background	1	Socioeconomic Status
Initial Achievement Status	1	Pretest
Educational Treatment	13	Utilization of Objectives
		Individualization in Decision Making
		Teacher or Locally Developed Materials
		Individualization of Instructional Pace
		Scheduling Characteristics
		Use of Performance Agreements
		Classroom Group Organization
		Teaching Unit Composition
		Completeness of Instructional Package
		Utilization of Student Evaluation
		Degree of Individualization*
		Level of Innovation*
		Number of Minutes per Day
Teacher Characteristics	3	Teacher Age
		Teacher Socioeconomic Background
		Teaching Qualifications

*Superindices based on the EdExAG scales in this table--see text for a discussion.

there may have been more than one adult in the classroom was not a teacher attribute but rather was a treatment characteristic; see the items for the EdExAG scale Teaching Unit Composition in Table III-11.)

The dependent variables related to these key predictor variables were the CTBS Reading Total, Language Total and Arithmetic Total Expanded Scale Scores. Each Student record in each analysis data set also contained a residualized posttest score that was equal to that student's posttest score minus the prediction of that score based on the regression of posttest on pretest and SES. The formula for these computations was:

$$Y' = Y - \hat{Y} \qquad \text{(IV-2)}$$

where,

Y' = the residual score;

Y = the original posttest score; and

$\hat{Y}$ = the posttest score predicted on the basis of the coefficients obtained in the regression of posttest on pretest and SES.

For any given analysis data set, the mean Y' score was, of course, equal to zero, and the variance was equal to $S_y^2(1 - R_{y.ps}^2)$, where S_y^2 was the variance of the posttest scores and $R_{y\cdot ps}$ was the multiple correlation between posttest and pretest and SES.

Other outcome measures, such as student attitudes, and other possible correlates of changes in school-related attitudes and/or in cognitive achievement (e.g., additional EdExAG or classroom observation scales) were considered in the analyses of unusually effective or ineffective EdExAG groups and consistently high-achieving or low-achieving students--see Questions 13 and 14.

7. How were student socioeconomic background and pretest confounded with the kinds of treatments and teachers to which students were exposed?

An inevitable occurrence in nonexperimental investigations conducted in public schools is that students with similar background characteristics tend to be found in the same schools. It would be expected, for example, that a neighborhood school would be fairly homogeneous with respect to student SES. Therefore, a major proportion of variation in SES would undoubtedly occur among schools and not within schools. Since EdExAG groups were largely synonymous with schools, especially at the elementary school level, EdExAG group membership and SES were expected to be confounded in our data.

In addition, it was also reasonable to expect that EdExAG group membership would be confounded with pretest. ("Ability grouping" on the basis of prior achievement, for example, would represent an extreme example of such confounding.) Although there were a number of probable reasons why this confounding could be expected to occur, the most important consideration in terms of analytic and interpretive consequences was that pretest undoubtedly reflected all those factors influencing achievement to which the student was exposed prior to his participation in Project LONGSTEP. One such factor could have been systematic exposure to unusually effective or unusually ineffective educational experiences across school years. More specifically, if the students in a given EdExAG group happened to have been exposed to, say, an unusually effective treatment prior to the study, as well as during the study, then part of the correlation between pretest and posttest would be due to the operation of a third variable, educational treatment. Attributing all of the pretest/posttest association to individual differences in achievement (i.e., to pretest) would, of course, result in an underestimation of the total impact of the treatment over time (though not of the impact during the year in question).

Thus, pretest was not considered a measure of pretreatment, even though it did provide a baseline measure for assessing gains in achievement during one school year. Unfortunately, there was no way to discover exactly why EdExAG groups differed on the pretest because there was no way to determine if the groups were equivalent prior to any schooling. In fact, variations among EdExAG groups with respect to SES backgrounds suggested that the students in these groups probably differed at the start of their schooling. Nevertheless, since pretest and SES were to be the primary covariates in our analysis of educational growth, it was important to assess the degree to which groups differing with respect to educational approach and with respect to the characteristics of their teachers also differed with respect to background and initial achievement status. These trends were analyzed by examining (a) the correlations among pretest, SES and the key EdExAG and teacher variables and (b) the commonality analyses in which pretest or SES was used as the dependent variable.

Commonality analysis is a technique that yields a series of coefficients which express the extent to which each independent variable (or each set of independent variables) in a multiple regression equation (a) is a unique, non-overlapping predictor of the criterion and (b) is a confounded predictor of the criterion due to its correlations with other independent variables in the model. Only a "uniqueness" may be viewed as reflecting a proportion of criterion variance since it is the square of the correlation of an independent variable with the criterion, after that independent variable has been residualized on the basis of all other independent variables in the model. This procedure has been described, as well as extensively used, by Mayeske and his colleagues (1972, 1973a, 1973b) in their reanalyses of the Equal Educational Opportunity Survey data (Coleman et al., 1966). Other authors (Mood, 1971; Kerlinger and Pedhazur, 1973; and Tatsuoka, 1973) have also discussed this method. A more complete discussion of commonality analysis and the manner in which it was used to analyze confounding among EdExAG-level and student-level variables is found in Appendix IV-E.

8. How was posttest related to the key analysis variables and to EdExAG group membership?

Before a multipredictor model of educational growth was analyzed, project staff examined the correlations of posttest with the other key analysis variables to see if these simple associations demonstrated any consistent evidence for the presence of substantial treatment effects. In addition, since the treatment variables utilized in these and in later analyses were collected at the EdExAG group level, it was valuable to assess (a) the proportion of posttest variance that could be attributed to EdExAG-group differences, regardless of the nature of the treatments present in these groups and (b) the extent to which pretest, SES, the teacher variables, and EdExAG-group membership were confounded or overlapping predictors of posttest performance.

The relationship of posttest performance to EdExAG-group membership and the degree to which group differences were confounded with pretest, SES and teacher variable differences was examined by means of regression analysis and commonality analysis in which EdExAG-group membership was represented by a set of dummy variables. As described more fully in Appendix IV-E, it can be shown that the multiple correlation (squared) of a dependent variable with a series of dummy variables encoding mutually exclusive group membership expresses (a) the ratio of among-group criterion variance to the total variance, and (b) the maximum proportion of criterion variance that can be attributed to any group-level variable or variables. (A group-level variable is a measure that has no intragroup variation and is constant for all members of the same group.) Since EdExAG-level scales are a kind of group-level variable, regressing a student-level dependent variable on the dummy variables encoding EdExAG group membership indexed the maximum proportion of criterion variance that could be associated with EdExAG-level variables.

The use of these dummy variables and SES, pretest and teacher variables as separate sets of predictors in a commonality analysis also permitted the assessment of the nature and extent to which these sets of variables were confounded predictors of posttest performance. As shown in Appendix IV-E, the use of the dummy variables encoding EdExAG-group membership in these commonality analyses allowed the within-groups confounding of pretest, SES and teacher variables to be distinguished from the among-groups confounding.

Following these analyses, the degree of relationship between specific treatment attributes of the EdExAG groups and posttest were examined. One issue that arose in these analyses was that of the most appropriate statistical test for the correlation of an EdExAG-level variable and a student-level dependent variable such as posttest. It was apparent that a simple t test of the correlation was not completely appropriate because an EdExAG variable was not really a concomitant of posttest performance, but was considered to be one of its possible antecedents. For this reason, it seemed most reasonable to test an EdExAG-variable/posttest relationship by means of a regression analysis approach to linear trend analysis. The specific formula utilized and some of the issues involved in its development are discussed in the following paragraphs.

Although the square of the EdExAG-variable/posttest correlation indicated the proportion of total variance in posttest that was linearly associated with the EdExAG variable, the remaining "unpredicted" posttest variance did not necessarily represent only prediction error. Rather, due to the structure of LONGSTEP treatment data, some of this variation could represent systematic differences among EdExAG groups and/or a non-linear effect of the treatment variable. Both possibilities were present because EdExAG groups could be nested under levels of a given EdExAG scale (i.e., more than one EdExAG group may have obtained the same score on a given EdExAG scale), either because the resolving power of the scale was less than that of the items differentiating one EdExAG group from another or because different schools were involved. (Table IV-8 is an example of how EdExAG groups were nested under levels of the Individualization of Instructional Pace scale for the Cohort 1, Spring 1972 Reading Total Score sample.)

TABLE IV-8

Posttest Means, Standard Deviations,
Numbers of EdExAG Groups and
Students at Each Level of the
Individualization of Instructional Pace (IIP) Scale
(Cohort 1, Spring 1972 Reading Total Score Analysis)

IIP Scale Score/Level	Mean	Standard Deviation	Number of EdExAG Groups	Number of Students
1.0	335.1	67.0	3	101
2.0	342.0	62.7	9	382
3.0	348.2	52.9	6	308

The nesting of EdExAG groups under different levels of a given EdExAG scale meant that criterion variation among EdExAG groups could be attributed to two sources: the variation among EdExAG groups *within levels* of the EdExAG variable (pooled across the various levels) and the variation *among levels* of the EdExAG variable. This latter source could be further partitioned into linear and nonlinear components. The sum of squares attributable to the linear component divided by the *total* sum of squares is, of course, the square of the criterion/EdExAG variable correlation. Figure IV-1 shows the sources to which variation in a student-level dependent variable could be attributed.

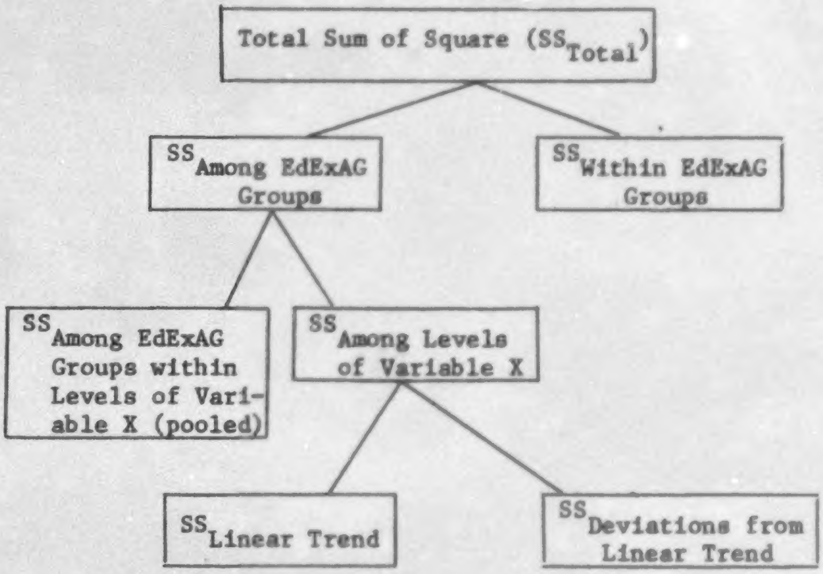

Total Sum of Square (SS_{Total})
$SS_{Among\ EdExAG\ Groups}$
$SS_{Within\ EdExAG\ Groups}$
$SS_{Among\ EdExAG\ Groups\ within\ Levels\ of\ Variable\ X\ (pooled)}$
$SS_{Among\ Levels\ of\ Variable\ X}$
$SS_{Linear\ Trend}$
$SS_{Deviations\ from\ Linear\ Trend}$

Figure IV-1. Variance sources for the regression of a student-level dependent variable on a quantitative EdExAG-level predictor (X).

Rather than assume that group differences and nonlinearities were not present, it seemed more reasonable to base the estimate of prediction error on the variation of students within EdExAG groups. Consequently, the following formula was used to test the linear relationship between an EdExAG variable and a student-level dependent variable:

$$F = \frac{r^2_{yx}}{(1 - R^2_{y \cdot D_j})/(N - n)} \; ; \qquad \text{(IV-3)}$$

where,

r^2_{yx} = the square of the correlation between a student-level dependent variable, Y, and an EdExAG-level variable, X ;

$R^2_{y \cdot D_j}$ = the square of the multiple correlation between Y and a series of dummy variables (D_j) that encode EdExAG group membership;

N = the total number of students; and

n = the number of EdExAG groups.

df = 1, (N - n).

Since the simple zero-order correlation of an EdExAG variable with posttest included variation in posttest expected on the basis of pretest and SES alone, the key EdExAG variables were also correlated with posttest residualized on the basis of pretest and SES. It should be emphasized that these correlations did not involve the same statistical model or the same educational model as that involved in the growth model (discussed under Question 9) or in the dummy variable prediction model discussed previously.

Multipredictor models of posttest performance ascertain the proportion of total variance in posttest that can be attributed to EdExAG variables (or EdExAG-group membership) residualized on the basis of pretest and SES--that is, the increase in a multiple correlation when a treatment variable is added to a prediction equation containing pretest and SES. The square of the correlation of posttest (Y) with that part of an EdExAG variable (X) that is independent of pretest (P) and SES (S) can be shown to be equal to:

$$r^2_{y(x \cdot ps)} = R^2_{y \cdot psx} - R^2_{y \cdot ps} \; ; \qquad \text{(IV-4)}$$

where,

$r^2_{y(x \cdot ps)}$ = the square of the correlation between Y and that part of X which is uncorrelated with pretest and SES;

$R^2_{y \cdot psx}$ = the square of the multiple correlation of Y with pretest, SES and EdExAG variable X; and

$R^2_{y \cdot ps}$ = the square of the multiple correlation of Y with pretest and SES.

The educational model implicit in this part correlation assumes that all posttest variance that is common to pretest, SES and the treatment variable should be attributed to pretest or SES effects. In short, the treatment variable is "adjusted" for pretest and SES differences and the adjusted score is related to achievement.

On the other hand, the correlation of an EdExAG variable with a residualized posttest score is not usually equal to $r_{y(x \cdot ps)}$ and can be shown to be equal to:

$$r^2_{x(y \cdot ps)} = r^2_{y(x \cdot ps)} \frac{1 - R^2_{x \cdot ps}}{1 - R^2_{y \cdot ps}} \qquad \text{(IV-5)}$$

where,

$r^2_{x(y \cdot ps)}$ = the square of the correlation between X and that part of posttest (Y) that is uncorrelated with pretest and SES;

$r^2_{y(x \cdot ps)}$ = the square of the correlation between posttest (Y) and that part of X that is uncorrelated with pretest and SES;

$R^2_{x \cdot ps}$ = the square of the multiple correlation between the EdExAG variable X with pretest and SES; and

$R^2_{y \cdot ps}$ = the square of the multiple correlation between posttest (Y) and pretest and SES.

The educational model implicit in this part correlation is that the treatment could have affected only the variation in student performance that was independent of that expected on the basis of the student's pretest and SES level. Although Werts and Linn (1970, p. 19) have suggested that such an approach can underestimate treatment effects when "other determinants of growth are correlated with initial status," this strategy was implemented so as to provide a slightly different methodological approach to the examination of treatment effects than was utilized later.

9. What growth model was examined?

Although the creation of a model of educational growth from correlational data is not fully justifiable, the examination of such a model did represent a meaningful way in which to approach the analysis of the LONGSTEP data base. "While the real world certainly does not really operate according to the model [or models] discussed, at least predictions can be made which further research can either confirm or disprove" (Linn, Werts and Tucker, 1971, p. 86). This approach was utilized for a number of reasons. First, as described in Chapter III, the development of the analysis variables themselves had been based primarily on a priori reasoning and the belief that such measures would relate meaningfully to public school educational approaches. It was certainly true that these judgments, as well as those involved in selecting certain variables for analysis, implicitly involved hypotheses as to the kinds of school factors related to longitudinal growth in achievement. Second, examination of such a model would serve to clarify the appropriateness of the statistical methodology used to analyze the LONGSTEP data. After all, "one cannot specify which statistics are appropriate until the (postulated) nature of the phenomena under study has been specified" (Linn, Werts and Tucker, 1971, p. 85). Third, constructing and testing a growth model with the LONGSTEP data would assist "in clarifying which data need to be collected in the next survey to help choose among the plausible alternative hypotheses for observed assocations" (Linn, Werts and Tucker, 1971, p. 86).

The initial growth model that was developed included the predictor variables shown previously in Table IV-7 (excluding the superindices, Degree of Individualization and Level of Innovation). Including these variables as predictors was a natural first step since each represented an hypothesized correlate of achievement. Also, the treatment variables were quantitative measures of many of the major ways in which LONGSTEP schools and EdExAG groups did differ from each other. This model, then, was comprised of a measure of the student's home background; a measure of initial achievement status; ten EdExAG-level variables quantifying some of the major attributes of each student's educational treatment; a measure of the amount of class time spent per day on the subject (language arts or math); and three aggregate teacher variables which described the qualifications, age and socioeconomic background of each student's language arts or math teacher(s). To test the feasibility of this approach, the initial growth model was first used to analyze the Spring 1972 Reading data for Cohort 4. Cohort 4 was used because it contained a larger number of EdExAG groups and a larger number of students than either Cohort 1 or Cohort 6.

Coefficients for the variables in the model were obtained by regressing the CTBS Reading Total Expanded Scale Score (corrected for site differences in proportion of school year completed prior to test administration) on the 16 predictors. Table IV-9 shows the zero-order correlation of each variable with the criterion; the raw score and standard score regression coefficients; the t test for each coefficient; and a measure of the extent to which the variable was confounded with other variables in the model (the square of the multiple correlation, $R^2[I]$, between each predictor and the remaining 15 predictor variables). Although the overall model was highly predictive of posttest performance, an examination of the zero-order correlations with posttest showed that pretest ($r = .84$) and SES ($r = .37$), as expected, were undoubtedly the primary reasons for the high multiple R. The only treatment variable receiving a statistically significant regression coefficient was Completeness of Instructional Package. The fact that the correlation of this variable with posttest was low and positive ($r = .04$), and that it was highly confounded with the other predictors in the model (see $R^2[I]$ in Table IV-9), strongly suggested that its statistically significant negative coefficient probably would not cross-validate.

TABLE IV-9

Multiple Regression Analysis of the Spring 1972 Reading Total Criterion -- Cohort 4 (N = 1952; n = 42)

Variable	r	B	β	t	R^2(I)
(Intercept)		23.948			
Reading Total for SP71 Residualized for Time of Testing	.838	.860	.805	60.998*	.152
Socioeconomic Status	.371	.647	.094	6.620*	.262
Utilization of Objectives	.045	2.998	.037	1.311	.812
Individualization in Decision Making	-.036	-3.283	-.020	1.092	.583
Teacher or Locally Developed Materials	-.024	2.043	.016	.917	.536
Individualization of Instructional Pace	.065	2.699	.026	1.227	.663
Scheduling Characteristics	-.066	-1.330	-.012	.684	.522
Use of Performance Agreements	.016	- .019	-.000	.010	.520
Classroom Group Organization	-.026	-3.197	-.034	1.543	.694
Teaching Unit Composition	-.033	.794	.008	.482	.486
Completeness of Instructional Package	.042	-7.419	-.065	2.706*	.743
Utilization of Student Evaluation	.057	- .213	-.002	.078	.782
Number of Minutes Per Day	.041	.062	.024	1.454	.470
Teacher's Age	.048	.156	.024	1.390	.522
Teacher's Socioeconomic Background	.072	.139	.020	1.277	.383
Teaching Qualifications	.053	- .105	-.012	.638	.583

R = .84530 R^2 = .71453 F = 302.71 df = 16, 1935 p<.0001

*p <.01 (two-tailed)

Notation:

N = number of students
n = number of EdExAG groups
r = zero-order correlation with posttest
B = raw score regression coefficient
β = standard score regression coefficient
t = t test of the regression coefficient (df = 1935)
R^2(I) = a measure of collinearity equal to the multiple correlation squared of the indicated predictor with all other predictor variables in the model

Further inspection of $R^2(I)$ in Table IV-9 shows that all of the EdExAG variables were highly confounded. Unfortunately, as numerous authors have pointed out (e.g., Darlington, 1968; Farrar and Glauber, 1967; Gordon, 1968), it becomes increasingly difficult to interpret the independent influence of a given variable when that variable is highly correlated with another variable or other variables in the regression model. In short, "when orthogonality is absent the concept of the contribution of an individual regressor remains inherently ambiguous" (Goldberger, 1964, p. 201).

The degree of multicollinearity shown in Table IV-9 was undoubtedly due in part to the fact that characteristics of educational treatments tend to occur together in public schools. The resulting intercorrelations among the EdExAG scales may have reflected the dependencies that did, in fact, exist. Also, since EdExAG scale scores were attached to all students in an EdExAG group, correlations among such variables were not based on the total number of students but on the number of EdExAG groups. The number of students in a given EdExAG group only weighted the contribution of one pair of scores to the overall correlation. Because multiple correlations approach 1.0 as the number of degrees of freedom shrinks, the multiple correlation between one EdExAG scale and the remaining EdExAG scales would tend to approach 1.0 if the number of EdExAG groups were small relative to the number of EdExAG scales in the model. This was probably a major reason why the EdExAG-level variables in the Spring 1972 Reading Total Score analysis sample for Cohort 4 were so highly confounded--there were only 42 EdExAG groups while there were 11 EdExAG-level variables in the model.

Appendix IV-F provides a summary of the number of schools and EdExAG groups per site present in each of the 18 analysis data sets discussed in this report. Examination of the appendix shows that the number of EdExAG groups would limit the applications of a growth model containing 11 EdExAG-level variables, especially for analyses involving Cohorts 1 and 6. Therefore, project staff decided that a more parsimonious growth model should be examined. The individual EdExAG variables would be used to describe the attributes of EdExAG groups whose performance appeared to be unusually effective (see Question 12). Substantial zero-order correlations of these treatment variables with posttest and with a posttest residual score were also noted (see Question 8).

The development of a revised growth model actually led to two models which were used to test two key hypotheses. The first model was designed to test the extent to which Level of Innovation (see Table IV-7) was related to growth in achievement, while the second model related Degree of Individualization (see Table IV-7) to growth. The Level of Innovation model contained the following variables as predictors: SES, Pretest, Level of Innovation, Number of Minutes per Day (spent on language arts or math) and Teaching Qualifications. The Degree of Individualization model was the same except that Degree of Individualization provided the primary measure of treatment in place of Level of Innovation.

Although the initial growth model contained three teacher variables, Teacher Socioeconomic Background, Teacher Age and Teaching Qualifications, only the last of these measures was included in the two revised growth models. This was done in the interest of parsimony and because of the magnitude of the correlations among these variables.

10. Did socioeconomic status or pretest act as a moderator variable in the growth model?

The regression of posttest on the predictor variables in the two revised growth models was based on the assumption that the impact of the variables was additive. In analysis of variance terms, this assumption meant that there were no interactions among the independent variables, and the relationship between the dependent variable and a particular independent variable did not depend on the level of another variable or combination of variables. It was possible to test this assumption directly by developing a regression model with interaction terms in it. This equation for the SES moderator analysis of the Level of Innovation model was:

$$
\begin{aligned}
Y = a_0 &+ a_1(\mathrm{SES}) + a_2(\mathrm{Pretest}) + \ldots \qquad \text{(IV-6)}\\
&+ a_3(\mathrm{LI}) + a_4(\mathrm{NMIN}) + \ldots\\
&+ a_5(\mathrm{TQ}) + a_6(\mathrm{Pretest} \times \mathrm{SES}) + \ldots\\
&+ a_7(\mathrm{LI} \times \mathrm{SES}) + a_8(\mathrm{NMIN} \times \mathrm{SES}) + \ldots\\
&+ a_9(\mathrm{TQ} \times \mathrm{SES}) + e \; ;
\end{aligned}
$$

where,

Y = posttest;

LI = Level of Innovation, the sum of the ten key measures of innovation;

NMIN = number of minutes per day spent on language arts or math (whichever is appropriate);

TQ = Teaching Qualifications;

a_0 = the raw score Y intercept;

a_1 to a_5 = the raw score regression coefficients for the growth model variables;

a_6 to a_9 = the raw score regression coefficients for the product (i.e., interaction) variables; and

e = a prediction error component.

After determining the multiple correlation squared for this full model (R^2_{FM}), restrictions were placed on the model that reflected the null hypothesis that there are no interaction effects. These restrictions were that the regression coefficients for the product variables were all equal to zero, or

$$a_6 = a_7 = a_8 = a_9 = 0.0 \; .$$

Substituting the restrictions into Equation IV-6 resulted in the product variables becoming equal to zero, leaving the restricted model,

$$
\begin{aligned}
Y = b_0 &+ b_1(\mathrm{SES}) + b_2(\mathrm{Pretest}) + \ldots \qquad \text{(IV-7)}\\
&+ b_3(\mathrm{LI}) + b_4(\mathrm{NMIN}) + \ldots\\
&+ b_5(\mathrm{TQ}) + e \; .
\end{aligned}
$$

The variables in Equation IV-7 were defined as in Equation IV-6, while regression coefficients b_0 to b_5 were those obtained in the regression of the dependent variable on just the five predictors themselves.

The multiple correlation squared for the restricted model (R^2_{RM}) was then computed and compared with that for the full model (R^2_{FM}) by means of an F test, in which

$$F = \frac{\left(R^2_{FM} - R^2_{RM}\right) \Big/ \left(NP_{FM} - NP_{RM}\right)}{\left(1 - R^2_{FM}\right) \Big/ \left[N - (NP_{FM} + 1)\right]} \; ; \qquad \text{(IV-8)}$$

and where,

R^2_{FM} = the square of the multiple correlation including the interaction (i.e., product) variables;

R^2_{RM} = the square of the multiple correlation including just the original predictor variables;

NP_{FM}, NP_{RM} = the number of predictors in the full and restricted models, respectively;

N = the total number of students; and

df = $(NP_{FM} - NP_{RM})$, $[N - (NP_{FM} + 1)]$.

Given the usual assumptions that the prediction error is normally distributed and that it is constant across the range of predicted scores (i.e., the assumption of homoscedasticity), the F test yielded by comparing each full model R^2_{FM} with the restricted model R^2_{RM} provided statistical evidence that at least one of the interactions tested was not equal to zero in the population of which the sample was representative.

Project staff considered the possibility that either of the two growth models might depend on the SES level or the pretest level of the students. Regression models including interaction components were developed and tested to see if there was reason to believe that Level of Innovation or Degree of Individualization were differentially effective for students at different SES or pretest levels.

It is important to note that because of the large numbers of students in these analyses and the high predictability of the models due to the presence of pretest and SES as predictors, it was expected that a number of statistically significant differences would be obtained. Project staff examined the results to see whether or not the magnitude of the moderator effect (indexed by the difference between the full model R^2_{FM} and restricted model R^2_{RM}) warranted the consideration of a substantially more complex analytic model. It was also necessary to interpret the findings of these analyses in light of the confounding that existed among SES or pretest and the treatment variables--that is, it was necessary to consider alternative explanations for the appearance of moderator effects, regardless of their magnitude.

Appendix IV-G contains these results for all analysis data sets discussed in this report. Although a number of the F tests were statistically significant ($p < .05$), most of the effects, as indexed by $R^2_{FM} - R^2_{RM}$, were very small. The additional criterion variance that was explained by adding the interaction variables to the model was usually very small (i.e., about 1 percent or less).

Pretest and SES interactions in both the Level of Innovation and the Degree of Individualization models were minute for all Reading Total Score analyses. Of the statistically significant interactions present in the Language Total and Arithmetic Total analyses, the largest seemed to have occurred in the data sets for Cohort 1, although even these effects were very small. Since there did not appear to be substantial interactions in any of the data sets, project staff decided not to include interaction components in the growth models to be examined in subsequent analyses.

11. What relationships existed among the predictor variables in the growth models and to what extent were these variables unique or overlapping predictors of posttest performance?

Some of the best documented results of experimental and non-experimental educational research efforts are the substantial positive correlations among posttest, pretest and student socioeconomic status. Since EdExAG group membership was correlated with SES and with pretest, and since SES, pretest and EdExAG-level variables were included in the revised growth models, it was expected that these variables would be confounded or overlapping predictors of posttest performance. A commonality analysis of the four categories of variables in each growth model--SES, pretest, the EdExAG variables and Teaching Qualifications-- was used to assess the degree and nature of this confounding. As noted previously with regard to Question 7, this technique has been used extensively with non-experimental school effects data (Mayeske et al., 1972; Mayeske, Okada, and Beaton, 1973; Mayeske, Okada, Beaton, Cohen, and Wisler, 1973). Appendix IV-E contains a summary of the method. This same appendix also describes how regression analysis and commonality analysis may be used with multilevel data so as to quantify (a) the maximum proportion of criterion variance attributable to any one or any combination of EdExAG-level predictor variables and (b) the predictive overlap among student-level and EdExAG-level predictor variables. A comparison of the regression analysis and commonality analysis results for the growth models (based on the EdExAG variables) with the results for the analyses based on the dummy variables encoding EdExAG group membership (described under Question 8) permitted an assessment of the degree to which the EdExAG-level variables in the model accounted for EdExAG-group differences that were present.

Again, it is important to note that EdExAG groups were not created specifically in response to variation with respect to the key analysis variables listed in Table IV-7. As discussed in Chapters II and III, different EdExAG groups were identified within a school when an AIR site visitor could differentiate among the treatments received by different groups of students by means of items on the EdExAG. Thus, two groups of students could belong to two different EdExAG groups because (a) two different schools were involved (EdExAG groups did not cut across schools), or (b) two groups within the same school could be differentiated on the basis of one or more EdExAG items. The variables shown in Table IV-7 are, in many instances, multi-item scales. Since different combinations of items could lead to the same scale score on a multi-item scale, it was possible that two EdExAG groups were identified within a school because of two discrepant items, but that the two groups had exactly the same scale score on the scale of which the items were members. It was also possible that groups were differentiated on the basis of items other than those involved in these key scales. In summary, EdExAG groups did not have to be different with respect to the scales examined for this report because the variables used to define these groups (i.e., the EdExAG items) made finer distinctions than did this select group of scales.

12. What were the regression coefficients for the variables in the two growth models and how were they interpreted?

After assessing the proportions of criterion variance that could be attributed to the various categories of indicators in each growth model (Question (11), the regression coefficients and the zero-order correlations among the variables in each analysis data set were examined and compared. Werts and Linn (1970, p. 18) have demonstrated algebraically that " when initial status is an independent variable, the regression coefficient for the other independent variable is the influence of that variable on growth." These same authors also make a number of other points that are important to remember regarding the interpretation of these results. For example, such models do not permit the prediction of school output from input "unless we know what the association of X [i.e., initial status] with Z [i.e., a treatment variable] means...." They state further:

> If the correlation between X and Z resulted from Z influencing X, then a change in Z will change X, which in turn will change Y [i.e., the dependent variable], in addition to any direct influence Z has on Y.... If the meaning of the correlation of X and Z is obscure then the researcher can talk only about "direct" influences, which may be a theoretical fiction since it may not be possible to ever measure this "direct" influence in isolation. If this "direct" influence is zero, one cannot assert that the variable does not influence the output since the influence may be exerted through effects on other variables which in turn influence output (pp. 21-22).

Because of the confounding that did exist among the predictors in the growth models, alternative explanations for the observed regression coefficients and proportions of criterion variance attributable to the various predictors (discussed under Question 11) were considered; these considerations were based on the trends demonstrated by those procedures described under Questions 7, 8 and 9.

13. When posttest was residualized on the basis of SES and pretest, were there any EdExAG groups whose mean residual scores were substantially different from zero? Could EdExAG groups which did better than expected (on the average) be differentiated consistently from those which did worse than expected?

Although the overall analyses conducted at earlier stages might have suggested that the treatment variables in the growth model had extremely small effects, it was possible that the students in a subset of EdExAG groups might, as a group, have done much better (or worse) than would be expected from their pretest and SES scores. For this reason, posttest was regressed on the two variables which, when taken together, are almost always found to be extremely good predictors of posttest performance: SES and pretest. Residual scores were then computed according to the formula:

$$Y' = Y - \hat{Y} ; \qquad \text{(IV-9)}$$

where,

Y' = posttest residual score (i.e., that part of the posttest that is uncorrelated with pretest and SES);

Y = posttest raw score; and

$\hat{Y}$ = posttest score predicted from pretest and SES.

Since the mean residual score was zero, students with positive residuals were viewed as doing better than "expected," while students with negative residual scores were viewed as doing worse than "expected." It should be noted that doing "better than" and "worse than" expected was relative to the performance of all the students with the same SES and pretest scores in a particular Project LONGSTEP analysis sample.

The mean residual score for all students in each EdExAG group was determined and the groups with the greatest positive and the greatest negative means were identified. Means and standard deviations for all variables in the growth model, as well as for most of the indices discussed in Chapter III, were then computed for the target groups and compared with each other by project staff. Since the number of variables was rather large, they are listed in Appendix IV-H.

14. Were unusual longitudinal growth patterns of individual students associated with certain student attitudes or characteristics, with certain educational practices or procedures, or with teachers with certain characteristics during the 1971-72 and 1972-73 school years?

Once a posttest score residualized for pretest and SES had been attached to each student record in the Spring 1972 and Spring 1973 analysis data sets, it was also possible to identify those individual students in each data set whose growth in achievement was greater or less than expected. And, when the two data sets were merged (on the basis of students who were present in both analysis samples), it was possible to identify students falling into one of the four categories shown in Figure IV-2.

		Year 1 to Year 2 Growth	
		Greater than Expected	Less than Expected
Year 2 to Year 3 Growth	Greater than Expected		
	Less than Expected		

Figure IV-2. Possible growth patterns of students who obtained posttest scores which deviated from those predicted from pretest and SES.

Outlying students were defined as those whose Spring 1972 and Spring 1973 residual scores were equal to or greater than (in absolute value) one-half of a standard deviation from the mean of the residuals. It was assumed that these cutoff scores permitted the identification of individual students who had demonstrated reliable and substantial deviations from expected achievement. Students exhibiting these four patterns of longitudinal achievement, particularly the consistent "overachievers" and "underachievers," were then compared and contrasted in terms of the profile variables shown in Appendix IV-H so that any correlates of the patterns could be identified.

Setting the cutoff score at ± one-half of a standard deviation should have minimized the possibility that students were classified as underachievers or overachievers simply because the test was not perfectly reliable. In any case, if misclassification due to unreliability were present to any substantial degree, it would serve to account primarily for students in the off-diagonal categories--i.e., students who overachieved during Year 2 and underachieved during Year 3, or vice versa. The performance of students who overachieved or underachieved both years, however, would have had to exceed ths apparent "regression-to-the-mean" effect. Therefore, it was reasonable to assume that these students legitimately represented a subset of students who demonstrated a reliable tendency to overachieve or underachieve for two consecutive school years.

V. RESULTS AND DISCUSSION

A. Introduction

The pretest/posttest analyses for each of three primary cognitive dependent variables and three analysis cohorts were conducted according to the analytic sequence presented in Chapter IV. The results of these analyses are presented in this chapter according to the particular dependent variable--reading, language or arithmetic/mathematics--that is examined. The results involving the CTBS Reading Total Score and the Language Total Score are presented in the first two sections, "The Analysis of Reading Achievement" and "The Analysis of Language Achievement." The language achievement analysis is discussed with reference to the reading achievement analysis since both the Reading and the Language Total Scores were related to the same language arts Educational Experience Analysis Guide (EdExAG) and teacher data. "The Analysis of Mathematics Achievement" is discussed in the section following reading and language.

Each of these three presentations of findings begins with a descriptive overview of the six data sets analyzed and an assessment of the extent to which the different cohorts (1, 4 and 6) and/or the different posttest analyses, Spring 1972 (SP72) and Spring 1973 (SP73), tended to involve substantially different kinds of treatments--at least different in terms of the key analysis variables examined for the report. Pretest and posttest means are also compared with the CTBS national norms published by CTB/McGraw-Hill so as to provide the reader with some idea of how students in the LONGSTEP sample performed relative to the CTBS (Forms Q and R) normative population.

Following the descriptive introduction, the analyses that focus on the extent to which treatment attributes were related to each other and to initial achievement status (pretest) and student socioeconomic background (SES) are reported. Commonality analyses with SES (or pretest) as the dependent variable and pretest (or SES), the EdExAG group dummy codes, and the three teacher variables as predictors are reported and discussed. These analyses focus on the extent to which pretest and SES were confounded with EdExAG group membership. Zero-order correlations are then examined to see if specific treatment attributes were confounded with each other, with student SES or with pretest.

The relationship of posttest scores with pretest, SES, the EdExAG dummy codes, and the three key teacher variables is discussed next. Commonality analysis results, as well as zero-order correlations, are reviewed. This discussion describes the manner in which socioeconomic background, initial achievement status, treatment group membership, and three teacher attributes were related to posttest performance and also the degree to which all predictors were confounded correlates of posttest achievement.

The results of the analysis of the two revised growth models are presented next--first the Level of Innovation model, which is the more general of the two, and then the Degree of Individualization model. Commonality analysis tables as well as standard score and raw score regression coefficients are shown, interpreted and discussed. Each of these two models represents a key hypothesis concerning the relationship between educational innovations and educational outcomes. The discussion describes whether or not these particular hypotheses are supported by the data and whether or not effects are substantial and/or consistent across school years and cohorts.

The last two sections discuss the attributes of (1) unusually effective and ineffective EdExAG/treatment groups, and (2) students who demonstrated substantial growth or lack of growth during two consecutive school years. These analyses were conducted because it was thought that substantial, consistent relationships might exist between treatment attributes and posttest for a small subset of EdExAG groups or individual students, even though such relationships may have been small when the total sample was analyzed.

The following chapter summarizes the trends that are described in detail in this chapter. Chapter VI also discusses implications and overall conclusions.

B. The Analysis of Reading Achievement

Descriptive Overview of the Analysis Samples

Pretest/posttest gains and national norms. Examination of the pretest and posttest means for the Reading Total Score analysis data sets (see Table V-1) reveals that pretest/posttest gains were not constant for all three analysis cohorts. In order to determine if any trends in gains were occurring across analysis samples, each pretest/posttest mean difference was expressed in average or pooled standard deviation units:

$$\frac{\overline{X}_{\text{posttest}} - \overline{X}_{\text{pretest}}}{SD_{\text{pooled}}}, \qquad \text{(V-1)}$$

where,

$$SD_{\text{pooled}} = \sqrt{[SD^2_{\text{posttest}} + SD^2_{\text{pretest}}]/2} \quad .$$

(Note: Sample size is not involved in this formula because the pretest and posttest distributions contained the same number of students.) Table V-2 shows the mean differences and "standardized" gains based on this equation. Since the greatest gains from one spring test administration to the next occurred in Cohort 1, while the smallest gains were found in Cohort 6, these results would seem to suggest that gains in reading achievement were negatively related to grade level in these LONGSTEP cohorts. The Expanded Scale Score national norms also tend to demonstrate this pattern of grade-level differences (see Table V-1). Although such differences in equal interval scale scores (which the Reading Total Expanded Scale Score was designed to be) may be due to test development or scaling procedures, it is also possible that they are reflecting the fact that younger students achieve proportionately greater gains in reading skills during one school year than do older students. Since such skills are just being developed during the early elementary grades, this latter conjecture is reasonable. However, the analytic implication of this pattern of change in achievement test scores is that it will be difficult to identify substantial treatment and/or teacher correlates of gains for older students who, as a group, demonstrate proportionately smaller amounts of growth during one school year.

TABLE V-1

Descriptive Summary of the Reading Total Score Analysis Samples

Variable	Cohort 1 SP72		Cohort 1 SP73		Cohort 4 SP72		Cohort 4 SP73		Cohort 6 SP72		Cohort 6 SP73	
Grade (pretest year/ posttest year)	1/2		2/3		4/5		5/6		6/7		7/8	
Total Number of Students	791		767		1952		1925		1520		1552	
Number of EdExAG Groups	18		17		42		44		11		13	
Number of Different Schools/Sites	13/5		13/5		35/13		34/13		7/7		7/7	
Mean ($\bar{X}$) and Standard Deviation (SD) by Variable	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD
Posttest-Reading Total Score Residualized for Time of Testing	339.98	63.33	400.66	66.24	469.51	68.67	507.40	76.14	531.27	77.72	563.43	83.92
National Norm - Posttest[1]	321.5	-	375.5	-	457	-	490.5	-	526	-	553.5	-
Pretest-Reading Total Score Residualized for Time of Testing	93.85[2]	10.56[2]	338.46	63.87	437.91	64.25	475.73	67.30	513.30	72.78	545.45	72.79
National Norm - Pretest[1]	85[3]	-	321.5	-	419	-	457	-	490.5	-	526	-
Socioeconomic Status	100.03	10.02	100.59	9.94	99.39	9.93	99.74	9.90	99.50	9.92	100.08	9.72
Utilization of Objectives	2.66	.72	2.68	.64	2.43	.84	2.48	.71	2.41	.76	2.47	.65
Individualization in Decision Making[4]	1.39	.53	1.50	.51	1.39	.43	1.39	.42	1.26	.30	1.25	.33
Teacher or Locally Developed Materials	1.65	.48	1.76	.37	1.81	.55	1.96	.57	2.34	.61	2.41	.49

(continued)

---

[1] The 50th percentile score in Expanded Scale Score units for testing conducted between month 6 and 9 for Form R (SP71,SP73) and Form Q (SP72) - from norms table found in the CTBS Examiner's Manual. Interpolated scale score norms have a decimal point.

[2] Based on CTMM Mental Age - Language score from Spring 1971.

[3] According to the CTMM Examiner's Manual a grade placement of 1.7 (the approximate time of testing) corresponds with a chronological age of 85 months.

[4] One of the scales summed to form the scale Degree of Individualization.

192

TABLE V-1 (continued)

Variable	Cohort 1 SP72		Cohort 1 SP73		Cohort 4 SP72		Cohort 4 SP73		Cohort 6 SP72		Cohort 6 SP73	
	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD	$\bar{X}$	SD
Individualization of Instructional Pace[4]	2.26	.67	2.23	.47	1.98	.65	2.19	.66	2.01	.76	1.93	.69
Scheduling Characteristics	1.72	.72	1.77	.63	1.59	.62	1.67	.59	1.79	.82	1.69	.65
Use of Performance Agreements[4]	1.63	.61	1.96	.85	1.50	.64	1.72	.77	1.44	.63	1.67	.76
Classroom Group Organization	2.12	.71	2.30	.46	1.87	.73	2.10	.62	2.03	.42	1.91	.35
Teaching Unit Composition	2.01	.71	2.21	.69	1.83	.71	1.84	.66	1.65	.60	1.70	.69
Completeness of Instructional Package	2.22	.64	2.11	.55	1.97	.60	2.05	.48	1.95	.62	1.90	.64
Utilization of Student Evaluation[4]	2.45	.62	2.51	.50	2.42	.65	2.55	.52	2.36	.73	2.32	.80
Degree of Individualization[5]	7.73	1.80	8.19	1.47	7.28	1.83	7.85	1.56	7.07	1.85	7.17	1.92
Level of Innovation[6]	20.12	4.33	21.03	3.26	18.77	4.43	19.95	3.55	19.24	4.74	19.25	4.46
Number of Minutes per Day	96.83	27.76	95.66	34.66	94.93	26.75	81.93	26.34	55.22	10.86	50.51	8.22
Teacher Age	31.91	7.80	35.37	10.49	36.50	10.75	36.18	9.96	33.31	9.91	37.16	9.78
Teacher Socioeconomic Background	102.54	7.87	101.70	10.45	98.92	9.74	98.78	10.18	105.78	7.87	102.45	9.78
Teaching Qualifications	97.61	7.84	100.90	6.35	102.47	7.88	99.86	7.46	95.26	8.24	99.06	6.53

---

[4] One of the scales summed to form the scale Degree of Individualization.

[5] A superindex equal to the sum of the four scales noted in footnote 4.

[6] A superindex equal to the sum of the ten EdExAG scales, Utilization of Objectives through Utilization of Student Evaluation.

TABLE V-2

Mean Differences between Reading Total Pretest and Posttest Scores and "Standardized" Gains

	Cohort 1[1] SP73	Cohort 4 SP72	Cohort 4 SP73	Cohort 6 SP72	Cohort 6 SP73
Mean Difference	62.20	31.60	31.67	17.97	17.98
"Standardized" Gain[2]	.96	.48	.44	.24	.23

[1] The SP72 comparison is not included for Cohort 1 because the CTMM Mental Age-Language was the pretest measure.

[2] From equation V-1.

Examination of Table V-1 and Figure V-1 also shows that the average student in these Project LONGSTEP analysis files scored slightly higher (but not substantially greater) than national norms. The difference between the LONGSTEP sample means and national norms for the posttest was slightly greater than the same difference for the pretest for Cohort 1, SP73. In contrast, for Cohorts 4 and 6, the posttest/norm differences were slightly lower than the corresponding pretest/norm differences. (The same trend was present when test/norm differences were expressed in test standard deviation units--see Appendix V-A.) Given the hypothesis that highly intensive, innovative educational programs should produce substantial gains in achievement, it was notable that the LONGSTEP sample posttest means were not conspicuously farther from their published norm than the pretest means were from their norm.

Two alternative explanations for these trends were considered:

- the possible effects of deleting students with missing data from the analysis samples, and
- the possible effects of deleting students with perfect or near perfect test scores from the analysis samples.

Deleting students with missing data from the analysis samples could have had some impact on the means because the occurrence of missing data

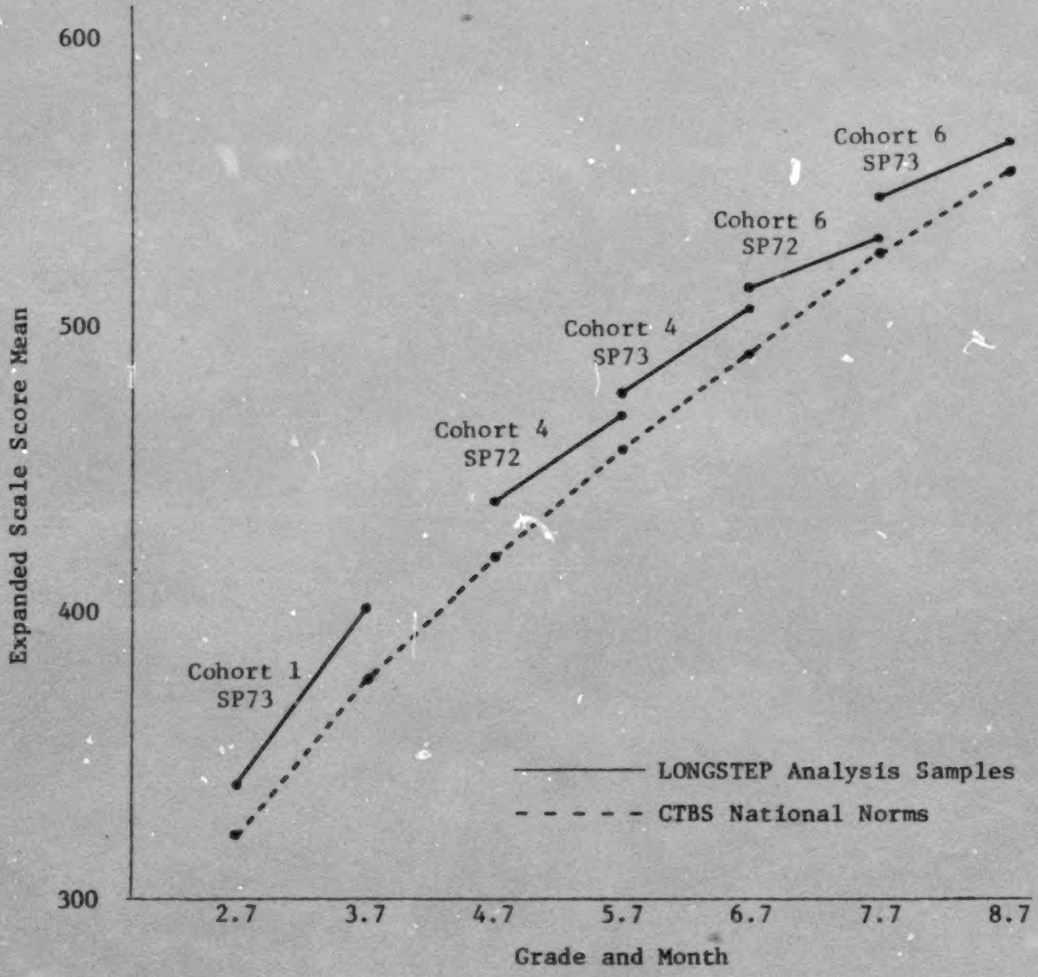


Figure V-1. Pretest and posttest means and published national norms (50th percentile for spring testing) - Reading Total Score analyses.

and level of achievement were inversely related (see the discussion of missing data in Chapter IV). Since attrition in the LONGSTEP sample was probably the primary reason for missing data, the trends shown in Figure V-1 may be reflecting selection effects due to student mobility. If it is assumed that growth rates at all levels of pretest are constant, deleting students with missing data probably had almost no differential effects with respect to the pretest and posttest means because both means would have been changed by a fairly constant amount due to the high correlation between pretest and posttest. On the other hand, it seems more likely that students with higher pretest scores had higher growth rates relative to lower scoring students. Thus, the gains shown in Figure V-1 are probably *greater* than would have resulted if the entire sample *could* have been analyzed because, as a group, students who had missing data had lower pretest scores (see Chapter IV). Slight selection or attrition effects of this nature do seem possible in Cohorts 4 and 6 because (1) the SP73 pretest mean is higher than the SP72 posttest mean for the *same* cohort (see Figure V-1) and (2) relative to their respective norms (see Appendix V-A), the losses are slightly smaller in the SP73 analysis samples than in the SP72 samples. Although such effects may have been present, and may have enhanced the gains shown by Cohorts 4 and 6, it is significant that neither pretest/posttest gains nor pretest and posttest norm comparisons suggested the presence of notable and consistent positive growth in achievement in these samples. Selection/attrition effects in the Cohort 1, SP73 analysis seem less likely because the SP72 sample posttest mean and SP73 sample pretest mean are almost identical (see Table V-1 and Figure V-1).

The possible impact of deleting from analysis those students who obtained perfect or near perfect test scores (i.e., those students who "topped out") was estimated by computing what the means would have been if these students had obtained the maximum pretest score *and* the maximum posttest score and if they had been included in the analysis samples (see Appendix V-B). Since students did not always top out on *both* pretest and posttest, this procedure would tend to overestimate the shift in pretest or posttest means that could be attributed to loss of data due to unavoidable test ceiling effects. As expected, because the number of students who topped out was fairly small (see Table IV-1 in Chapter IV), the means were changed very little (see Appendix V-B). Thus, the deletion of students with perfect or near perfect test scores presumably had little differential effect on the trends shown in Figure V-1.

It is most reasonable to conclude, therefore, that when pretest, posttest and national norms are considered:

- consistent and substantial average growth in reading skills did not occur in these LONGSTEP samples,
- the greatest gains in reading achievement for these three LONGSTEP samples occurred in Cohort 1, and
- the magnitude of pretest-to-posstest gain tended to be inversely related to grade level.

Treatment attributes. Table V-1 also shows the mean and standard deviation of each key variable for each analysis sample. Inspection of the means shows that the three cohorts did not differ greatly with respect to any of the key analysis variables except Number of Minutes per Day (that is, amount of time devoted to language arts activities). As expected, seventh-graders (Cohort 6, SP72) and eighth-graders (Cohort 6, SP73) in the LONGSTEP sample spent substantially less class time per day on language arts than did students in lower grades.

Since the highest possible score on any of the EdExAG/treatment scales was 3.0, the lowest, 1.0, and the midpoint, 2.0, the means shown in Table V-1 suggest that, on the average, the students in the three cohorts were involved in school programs that were moderately high with respect to Utilization of (Behavioral) Objectives and Utilization of Student Evaluation but below the arbitrary midpoint of 2.0 with respect to Individualization in Decision Making. Close examination of Table V-1 suggests, however, that the similarities and differences among the various samples are really variable-specific. Many of the EdExAG scale means are fairly close to the theoretical midpoint (2.0) of the scale although there are quite a few exceptions. Also, cohorts tended to differ slightly with respect to the treatment scales, one of the most notable differences being a greater use of Teacher or Locally Developed Materials for Cohort 6. In addition, the differences between the mean treatment attributes to which a given cohort of students was exposed during Year 2 and during Year 3 do not appear to be large and systematic.

Relationships among Treatments, Pretest and SES

The two primary student background variables associated with educational achievement level for which data were collected by Project LONGSTEP were initial achievement status (measured by a pretest administered toward the end of the previous school year) and the socioeconomic level of the student's home environment (SES). It was important to assess the extent to which groups of students who were exposed to different educational approaches also tended to be different with respect to pretest and/or SES since EdExAG/treatment groups that differed on mean pretest and/or SES level would be expected to differ on posttest, even if the treatments had no differential effects at all. Commonality analysis, as well as the zero-order correlations of the key EdExAG scales with pretest and with SES, was used to examine these differences.

The commonality analyses of SES involved regressing SES on three sets of predictors: (1) pretest; (2) a series of dummy variables encoding EdExAG group membership; and (3) the three key teacher variables, Teacher Age, Teacher Socioeconomic Background, and Teaching Qualifications. Conversely, the commonality analyses of pretest involved regressing pretest on SES, the EdExAG dummy codes and the three teacher variables. The commonality tables for the SES analysis for each of the six analysis samples have been placed in Appendix V-C; the tables for the pretest analyses are found in Appendix V-D.

Table V-3 summarizes for each of the six analysis samples (1) the proportion of among-group variation in SES [$R^2_{SES}(D_j)$] and in pretest [$R^2_{pretest}(D_j)$], and (2) the proportion of variance in pretest that could be attributed uniquely to EdExAG group membership [$U_{pretest}(D_j)$] and to SES variation within EdExAG groups [$U_{pretest}(SES)$]. Examination of these results shows that major differences did exist among EdExAG groups with respect to SES level. For example, Table V-3 shows that 52 percent of the variance in SES scores for the Cohort 4, SP72 sample was among-group variance, or conversely, that only 48 percent of the total variance in student

TABLE V-3

Proportions of SES and Pretest Variance Attributable to EdExAG Group Membership

	Cohort 1 SP72	Cohort 1 SP73	Cohort 4 SP72	Cohort 4 SP73	Cohort 6 SP72	Cohort 6 SP73
SES[1]						
$R^2_{SES}(D_j)$	.37	.42	.52	.51	.30	.31
Pretest						
$R^2_{pretest}(D_j)$	.07	.19	.14	.17	.09	.07
$U_{pretest}(D_j)$[2]	.10	.08	.07	.07	.03	.01
$U_{pretest}(SES)$[3]	.06	.06	.04	.04	.05	.07

[1] Among-group differences with respect to SES that were independent of concomitant differences in pretest were not computed because of the theoretical primacy of SES; that is, it was reasonable to view SES as an antecedent of test performance but not vice versa.

[2] This commonality analysis uniqueness is equal to the proportion of variance in pretest that is attributable to EdExAG group membership after group differences with respect to SES and the teacher variables are taken into account.

[3] This uniqueness is equal to the proportion of variance in pretest that is attributable to SES, independent of group differences on SES and all teacher-variable differences; that is, it is a within-group uniqueness.

SES was within-group variance. Such results, however, are not particularly surprising since neighborhood schools are usually somewhat homogeneous with respect to SES and since EdExAG groups were, in many instances, synonymous with schools. The commonality analyses of SES (see Appendix V-C) also showed that the associations between student SES and the three teacher variables (Age, SES Background and Teaching Qualifications) were small and consistently present in each of the analysis samples, but that such associations were almost completely confounded with EdExAG group membership and pretest.

EdExAG group differences with respect to initial achievement level (i.e., pretest) were also present but were smaller than group differences in SES. In fact, after concomitant EdExAG group differences with respect to SES (and the three teacher variables) are taken into account, group differences with respect to pretest level decrease substantially in all but the Cohort 1, SP72 analysis sample [see $U_{pretest}(D_j)$ in Table V-3]. Interestingly enough, however, once differences among groups and teachers have been taken into account, the magnitude of the within-group relationship between SES and pretest is approximately the same in all six analysis samples [see $U_{pretest}(SES)$ in Table V-3].

The pattern of associations shown in Table V-3, however, seems to differ between the Cohort 6 samples and the Cohorts 1 and 4 samples. For example, although one might expect home background (as measured by SES) to have accounted for slightly less within-group variance in pretest for older students (Cohort 6) than for younger students (Cohorts 1 and 4), there was relatively similar within-group overlap between pretest and SES in all samples [see $U_{pretest}(SES)$ in Table V-3]. One reason for this may have been that SES variation within the EdExAG groups present in the Cohort 6 samples was proportionately larger than the SES variation within the Cohort 1 and Cohort 4 EdExAG groups. Subtracting the proportion of among-group variance in SES [$R^2_{SES}(D_j)$ in Table V-3] from unity yields the proportion of SES variance occurring within EdExAG groups and shows that the students within the Cohort 6 EdExAG groups were relatively more heterogeneous with respect to SES background than were students in the Cohort 1 and Cohort 4 EdExAG groups. This result was, in turn, probably due to the fact that, because of its neighborhood location, the student body of the typical elementary school is usually more homogeneous with respect to SES than the student body of a central middle school or junior high school. Cohort 6 students were of junior high age during the 1971-72 school year (grade 7) and during 1972-73 school year (grade 8).

A second issue prompted by Table V-3 is based on the findings that (1) Cohort 6 EdExAG group differences with respect to both pretest and SES were smaller (relative to the particular total variation in each sample) than were EdExAG group differences in Cohorts 1 and 4 and (2), as noted previously, the magnitude of the within-group pretest/SES relationship is similar across all samples. The question arises, "Relative to Cohorts 1 and 4, why do almost all of the EdExAG group differences in pretest [see $R^2_{pretest}(D_j)$ in Table V-3] for the Cohort 6 samples disappear when SES differences are accounted for [see $U_{pretest}(D_j)$ in Table V-3], especially when Cohort 6 EdExAG groups were the least different with respect to mean SES [see $R^2_{SES}(D_j)$ in Table V-3]?" These trends, like all the others in Table V-3, are not dramatic. Nevertheless, a possible explanation for the fact that SES differences account for a proportionately larger share of the pretest differences among Cohort 6 EdExAG groups than for the Cohort 1 and 4 groups is that (1) more ability grouping was taking place in the LONGSTEP elementary schools and (2) when multiple EdExAG groups were identified within an elementary school (see Appendix IV-F, Tables 1 through 6), there was a greater tendency for such groups to have been formed on the basis of prior achievement (relative to the Cohort 6 or junior high EdExAG groups).

The teacher variables were, as a group, less highly related to student pretest level than they were to SES. The major portion of the associations between teacher variables and pretest tended to be confounded with EdExAG group membership and SES [see the third-order commonality coefficients, C(S,D,T) in Appendix V-D], except in the case of Cohort 4, Spring 1972.

The correlations of SES and pretest with each other and with the key treatment and teacher variables showed that student SES and pretest level were associated beyond the chance level ($p < .05$) with the majority of the treatment and teacher scales. Table V-4 and Table V-5 contain the SES and pretest correlations for all six analysis samples.

As shown in Table V-4, in relation to students with higher SES, those with lower SES in each of these LONGSTEP analysis samples tended to be members of programs characterized by a

- lower Degree of Individualization,
- lower Level of Innovation, and
- greater Number of Minutes per Day (spent on language arts instruction).

Students with lower SES also tended to have a teacher or teachers whose SES background was low relative to the other teachers in the LONGSTEP sample and whose age was slightly above the LONGSTEP average. In the analysis samples for Cohort 1 and Cohort 6, there was a tendency for most of the significant correlations to be positive in both the SP72 and SP73 data sets for a given cohort. However, as shown in Table V-4, the trend for Cohort 4 was for the SP72 correlations with SES to be positive and the SP73 correlations to be negative. Since the SES measure utilized in these analyses was constant across time points for a given student, this pattern indicates that relative to all of the students in each Cohort 4 analysis sample, lower SES students were involved in treatments that were below average with respect to many of these key treatment attributes during the 1971-72 school year, but were above average in innovativeness in 1972-73.

The treatment and teacher variables were also correlated with student initial achievement level (i.e., pretest). These correlations are shown in Table V-5. As with SES, it is difficult to discern systematic trends in these correlations across cohorts. There was a tendency for a consistently positive relationship to exist between initial achievement level (i.e., pretest) and exposure to a specific treatment attribute in the 1971-72 and in the 1972-73 school years for students in Cohort 6. This consistency was less apparent in Cohorts 1 and 4. There was, however, a slight trend in Cohort 1 for pretest to be negatively related to these key treatment attributes in the SP72 data set and positively related to the same treatment attribute the next year (SP73). The pretest measure for

Table V-4

Correlations of SES with the Key Analysis Variables - Reading Total Score Analysis

Level of Correlation	Variable	Cohort 1 SP72	Cohort 1 SP73	Cohort 4 SP72	Cohort 4 SP73	Cohort 6 SP72	Cohort 6 SP73
Student/ Student	Pretest	.168	.405	.340	.375	.320	.356
	Teacher Age	.022	-.103	-.036	-.118	-.081	-.203
	Teacher Socio-economic Background	-.002	.036	.210	.330	.162	.075
	Teaching Qualifications	.144	.168	-.061	-.065	-.241	-.234
EdExAG/ Student	Utilization of Objectives	.309	.299	.290	.162	.192	.292
	Individualization in Decision Making	-.019	.043	-.005	-.079	.09[illegible]	.189
	Teacher or Locally Developed Materials	.033	-.191	.073	-.168	.004	-.060
	Individualization of Instructional Pace	.130	.236	.250	-.076	.030	.039
	Scheduling Characteristics	-.030	-.114	.055	-.113	.349	.368
	Use of Performance Agreements	-.041	-.095	.070	.087	.122	.369
	Classroom Group Organization	-.157	-.106	.000	-.227	.327	.027
	Teaching Unit Composition	.241	.191	.081	-.014	-.141	.072
	Completeness of Instructional Package	.219	.241	.312	-.032	.148	.160

(continued)

TABLE V-4 (continued)

Level of Correlation	Variable	Cohort 1 SP72	Cohort 1 SP73	Cohort 4 SP72	Cohort 4 SP73	Cohort 6 SP72	Cohort 6 SP73
	Utilization of Student Evaluation	.230	.275	.214	-.142	.261	.240
	Degree of Individualization	.108	.129	.189	-.057	.172	.293
	Level of Innovation	.142	.138	.206	.085	.190	.252
	Number of Minutes per Day	-.145	.174	.077	-.098	-.271	-.383
r ($p < .05$) - Student/Student Level[1]		.069	.070	.044	.044	.050	.049
r ($p < .05$) - EdExAG/Student Level[2]		.055	.054	.031	.031	.042	.041

Note: For convenience, all coefficients in this table were truncated to three decimal places--see Appendix V-E for correlations with five-place accuracy.

[1] Absolute value of the coefficient required for a correlation between two student-level variables to be statistically significant ($\alpha \leqq .05$, two-tailed). Due to the large number of degrees of freedom, the critical ratio $z = r/\sigma_r$ where $\sigma_r = 1/\sqrt{N}$ was used.

[2] Absolute value of the coefficient required for a correlation between one student-level variable (SES) and an EdExAG-level variable to be statistically significant ($\alpha \leqq .05$, two-tailed)--see equation IV-3 for the computing formula.

TABLE V-5

Correlations of Pretest with the Key Analysis Variables - Reading Total Score Analysis

Level of Correlation	Variable	Cohort 1 SP72	Cohort 1 SP73	Cohort 4 SP72	Cohort 4 SP73	Cohort 6 SP72	Cohort 6 SP73
Student/ Student	Socioeconomic Background	.168	.405	.340	.375	.320	.356
	Teacher Age	.076	-.016	.057	.026	-.103	-.157
	Teacher Socioeconomic Background	.014	-.032	.037	.157	.099	.060
	Teaching Qualifications	.038	.154	.083	-.001	-.174	-.148
EdExAG/ Student	Utilization of Objectives	-.063	.155	.020	-.030	.108	.134
	Individualization in Decision Making	-.059	.034	-.011	-.075	.056	.092
	Teacher or Locally Developed Materials	-.105	-.126	-.037	-.123	.103	-.024
	Individualization of Instructional Pace	-.009	.126	.055	-.102	.042	.003
	Scheduling Characteristics	-.060	-.094	-.075	-.146	.185	.158
	Use of Performance Agreements	-.089	-.021	.040	-.002	.166	.148
	Classroom Group Organization	.073	.002	-.024	-.184	.217	.019
	Teaching Unit Composition	-.119	.109	-.031	-.118	-.050	.029

(continued)

TABLE V-5 (continued)

Level of Correlation	Variable	Cohort 1 SP72	Cohort 1 SP73	Cohort 4 SP72	Cohort 4 SP73	Cohort 6 SP72	Cohort 6 SP73
	Completeness of Instructional Package	-.128	.056	.038	-.065	.083	.066
	Utilization of Student Evaluation	-.122	.164	.049	-.133	.124	.108
	Degree of Individualization	-.093	.095	.049	-.109	.132	.121
	Level of Innovation	-.097	.073	.005	-.161	.138	.107
	Number of Minutes per Day	-.036	.089	.000	-.046	-.173	-.170
r ($p < .05$) - Student/Student Level[1]		.069	.070	.044	.044	.050	.049
r ($p < .05$) - EdExAG/Student Level[2]		.018	.031	.016	.018	.014	.012

Note: For convenience, all coefficients in this table were truncated to three decimal places--see Appendix V-E for correlations with five-place accuracy.

[1] Absolute value of the coefficient required for a correlation between two student-level variables to be statistically significant ($\alpha \leqq .05$, two-tailed). Due to the large number of degrees of freedom, the critical ratio $z = r/\sigma_r$ where $\sigma_r = 1/\sqrt{N}$ was used.

[2] Absolute value of the coefficient required for a correlation between one student-level variable (Pretest) and an EdExAG-level variable to be statistically significant ($\alpha \leqq .05$, two-tailed)--see equation IV-3 for the computing formula.

the SP72 analysis for Cohort 1 was the CTMM Mental Age-Language score, while the CTBS Reading Total Expanded Scale Score was the SP73 analysis pretest measure. However, since it does not seem likely that a change in tests would result in a shift in the sign of these correlations, it must be concluded that, on the average, Cohort 1 students who were below average in reading achievement tended to be members of more innovative treatments during the 1971-72 school year than similar students were during the 1972-73 school year. ("Average," of course, refers to a mean based on the students in a given LONGSTEP analysis sample.)

The general lack of similar correlations between pretest and Project LONGSTEP's key treatment variables in two consecutive school years prompted speculations as to the possible reasons for such "consistent inconsistency." First, it is possible that the treatments remained fairly stable across two school years but that from one school year to the next, schools used different criteria to assign students to treatments within schools. The fact that EdExAG groups were largely synonymous with schools (see Appendix IV-F) and that the SES/treatment attribute correlations showed the same inconsistency suggests that this explanation is not entirely appropriate.

Second, it is possible that the EdExAG scales were longitudinally unstable indices of the constructs they purported to measure. This is unlikely, however, since the same project personnel visited the same schools in both years and since treatment documentation in 1972-73 was based upon noting changes from the previous years. It is more likely that the 1972-73 EdExAG data underestimated changes (and were more stable than they should have been).

Last, it is probably most likely that treatment changes implemented by the schools, as well as treatment impact during the pretest year, combined in unknown ways across this set of diverse schools to yield these dissimilar pretest/treatment variable correlations.

Inspection of the intercorrelations among the EdExAG and teacher scales themselves (see the correlation matrices in Appendix V-E) shows that the treatment scores were indeed highly intercorrelated. As noted in Chapter IV and in Appendix IV-E, however, these correlations are EdExAG-level correlations in which each one of the k scores or each pair of k scores (k = the number of EdExAG groups) has been weighted by the number of students in that particular EdExAG group. And, since the number of students in each EdExAG group varied considerably, two different pairs of scores could have had differential effects on the overall correlation (because the groups contained different numbers of students). An "effective sample size" which, to some extent, takes into account the differential weighting of sampling errors in such data was derived (see Appendix IV-E) and provides a more adequate approximation to the sample size to be used when testing the statistical significance of these correlations. Since the number of EdExAG groups is small, especially in relation to the number of students in each analysis sample, and since the sampling error of r is large when the sample size is small, many of the moderately large correlations among the EdExAG scales shown in Appendix V-E were not statistically significant (at $p < .05$, two-tailed). The correlations of the EdExAG variables with the two superindices, Degree of Individualization and Level of Innovation, were generally high. However, these correlations were not tested for significance in view of the part-whole effect (the superindices were computed from the same individual EdExAG scales with which they were correlated here). Perhaps the most noticeable trend in the EdExAG-level correlations for all of the analysis samples was that the correlations between Number of Minutes per Day, the study's measure of treatment exposure, and the treatment attributes measured by the key EdExAG variables were very small.

These analyses suggest that, as expected, the measures of student, treatment and teacher attributes obtained by Project LONGSTEP were interrelated with one another. The highly innovative nature of the LONGSTEP sample undoubtedly had some effect on these correlations, especially those involving the treatment variables. Nevertheless, these findings do suggest that schools with different kinds of students tended to be doing slightly different kinds of things, or perhaps that they were doing the same kinds of things but to a lesser or greater degree. Furthermore, since one of the criteria involved in the identification of the project's treatment scales was that they be somewhat conceptually distinct, these results would imply that, on the average, certain treatment components did tend to occur together in the LONGSTEP sample schools. Treatment components, however, were not perfectly correlated with each other, Thus, these results support the viewpoint that educational programs are composed of combinations of component processes and that even in highly innovative schools the presence of one component does not necessarily ensure that another will also be present. (The Educational Experience Analysis Guide was, for this very reason, designed to document the underlying attributes of the educational treatments to which participating students were exposed, independent of the program label--see Chapter II.)

Relationships of Key Analysis Variables with Posttest Reading Performance

Before examining the correlations between the key analysis variables and posttest, each posttest measure was subjected to a commonality and levels analysis so that the predictive relationships in a four-predictor model containing SES, pretest, EdExAG/treatment group membership and the three teacher scales (Age, SES Background and Teaching Qualifications) could be assessed. The following section contains the results of these computations for the six analysis samples examined for this report. After this section is a summary of the results for those readers who would rather not be burdened with technical details.

Technical discussion of EdExAg group commonality analyses. For those readers who are unfamiliar with the general format of the commonality analysis tables used in this report, Appendix V-F provides an overview of how these tables may be read. The following paragraphs cover more specifically the properties of commonality analysis as it was used here to analyze among-group confounding. Technical issues regarding the use of commonality analysis with group-level predictors (such as a set of dummy variables encoding mutually-exclusive group membership) are discussed in Appendix IV-E.

As noted in Chapter IV and explained in greater detail in Appendix IV-E, the square of the multiple correlation of a dependent variable, such as posttest, with a series of dummy variables encoding EdExAG group membership quantifies the proportion of variation in that dependent variable that is occurring among groups. For example, the value .13822 at the bottom of the third column of numbers in Table V-6 means that approximately 14 percent of the variation in posttest reading performance is occurring among EdExAG groups. Another way of interpreting this value is that it is possible to increase the accuracy with which posttest reading performance can be predicted in these data by about 14 percent by merely knowing the EdExAG group to which each student belongs. This value also implies that, on the average, mean differences among EdExAG groups with respect to posttest performance were of this magnitude (i.e., ≈ 14 percent of the total variation was attributable to mean differences). Each of the squared multiple correlations at the bottom of Tables V-6 through V-11 [i.e., the $R^2(X_i)$] expresses the proportions of variation in the reading posttest associated with variation in those particular sets of predictor variables.

The first-order commonality coefficients or uniquenesses [$U(X_i)$] show the proportion of variance in posttest that can be predicted by a given variable or variable set after that variable has been adjusted on the basis of its correlations with the other predictor variables. For example, the value .06757 at the top of the third column of numbers in Table V-6 means that about 7 percent of the variation among student reading scores was uniquely associated with EdExAG group membership after SES, pretest and teacher variable differences had been taken into account. The unique contributions of SES, pretest and the teacher variables are shown in the same row. However, because one of the sets of predictors is composed of the dummy variables encoding EdExAG group membership, these uniquenesses actually quantify the extent to which pretest, SES or the teacher variables are uniquely related to posttest performance within EdExAG groups (see Appendix IV-E for a more detailed explanation). Thus, Table V-6 shows that about 5 percent of the variance in the criterion or dependent variable, posttest, could be uniquely related to SES variation within EdExAG groups in that analysis sample.

The second-order [e.g., C(S,P)], third-order [e.g., C(S,P,D)], and fourth-order [i.e., C(S,P,D,T)] commonality coefficients are each relative

TABLE V-6

Commonality Coefficients* for the Regression of Posttest on SES (S), Pretest (P), EdExAG Dummy Codes (D) and Teacher Variables (T) -- Cohort 1, Spring 1972, Reading Total Score Analysis

	SES(S)	Pretest (P)	EdExAG Dummy Codes (D)	Teacher Variables(T)
$U(X_i)$	.05144 (.15717)	.08709 (.26607)	.06757 (.20643)	.00320 (.00978)
C(S,P)	.04731 (.14454)	.04731 (.14454)		
C(S,D)	.06978 (.21319)		.06978 (.21319)	
C(S,T)	.00032 (.00098)			.00032 (.00098)
C(P,D)		.00547 (.01672)	.00547 (.01672)	
C(P,T)		-.00064 (-.00195)		-.00064 (-.00195)
C(D,T)			.00109 (.00335)	.00109 (.00335)
C(S,P,D)	-.00319 (-.00975)	-.00319 (-.00975)	-.00319 (-.00975)	
C(S,P,T)	.00036 (.00110)	.00036 (.00110)		.00036 (.00110)
C(S,D,T)	-.00010 (-.00030)		-.00010 (-.00030)	-.00010 (-.00030)
C(P,D,T)		-.00076 (-.00232)	-.00076 (-.00232)	-.00076 (-.00232)
C(S,P,D,T)	-.00164 (-.00501)	-.00164 (-.00501)	-.00164 (-.00501)	-.00164 (-.00501)
$R^2(X_i)$	.16429	.13400	.13822	.00184

$R^2(S,P,D,T) = .32731$

*Below each commonality coefficient and in parentheses is the normalized coefficient; that is, the commonality coefficient divided by $R^2(S,P,D,T)$.

TABLE V-7

Commonality Coefficients* for the Regression of Posttest on SES (S), Pretest (P), EdExAG Dummy Codes (D) and Teacher Variables (T) -- Cohort 1, Spring 1973, Reading Total Score Analysis

	SES(S)	Pretest (P)	EdExAg Dummy Codes (D)	Teacher Variables(T)
$U(X_i)$	.00945 (.01429)	.39745 (.60097)	.13347 (.20181)	.00309 (.00467)
C(S,P)	.06730 (.10176)	.06730 (.10176)		
C(S,D)	-.00886 (-.01340)		-.00886 (-.01340)	
C(S,T)	-.00076 (-.00115)			-.00076 (-.00115)
C(P,D)		.02671 (.04039)	.02671 (.04039)	
C(P,T)		.01306 (.01975)		.01306 (.01975)
C(D,T)			.00715 (.01081)	.00715 (.01081)
C(S,P,D)	-.01058 (-.01600)	-.01058 (-.01600)	-.01058 (-.01600)	
C(S,P,T)	-.00162 (-.00245)	-.00162 (-.00245)		-.00162 (-.00245)
C(S,D,T)	.00055 (.00084)		.00055 (.00084)	.00055 (.00084)
C(P,D,T)		.00559 (.00844)	.00559 (.00844)	.00559 (.00844)
C(S,P,D,T)	.01935 (.02926)	.01935 (.02926)	.01935 (.02926)	.01935 (.02926)
$R^2(X_i)$	.07484	.51727	.17338	.04642

$R^2(S,P,D,T) = .66135$

---

* Below each commonality coefficient and in parentheses is the normalized coefficient; that is, the commonality coefficient divided by $R^2(S,P,D,T)$.

TABLE V-8

Commonality Coefficients* for the Regression of Posttest on SES (S), Pretest (P), EdExAG Dummy Codes (D) and Teacher Variables (T) -- Cohort 4, Spring 1972, Reading Total Score Analysis

	SES(S)	Pretest (P)	EdExAG Dummy Codes (D)	Teacher Variables(T)
$U(X_i)$	.00281 (.00384)	.50689 (.69233)	.02024 (.02765)	.00046 (.00063)
C(S,P)	.04535 (.06193)	.04535 (.06193)		
C(S,D)	.00419 (.00573)		.00419 (.00573)	
C(S,T)	.00001 (.00001)			.00001 (.00001)
C(P,D)		.06015 (.08215)	.06015 (.08215)	
C(P,T)		.01298 (.01773)		.01298 (.01773)
C(D,T)			.00035 (.00048)	.00035 (.00048)
C(S,P,D)	.07662 (.10465)	.07662 (.10465)	.07662 (.10465)	
C(S,P,T)	.00212 (.00289)	.00212 (.00289)		.00212 (.00289)
C(S,D,T)	.00129 (.00176)		.00129 (.00176)	.00129 (.00176)
C(P,D,T)		-.00654 (-.00893)	-.00654 (-.00893)	-.00654 (-.00893)
C(S,P,D,T)	.00523 (.00714)	.00523 (.00714)	.00523 (.00714)	.00523 (.00714)
$R^2(X_i)$	.13762	.70279	.16154	.01591

$R^2(S,P,D,T) = .73216$

* Below each commonality coefficient and in parentheses is the normalized coefficient; that is, the commonality coefficient divided by $R^2(S,P,D,T)$.

TABLE V-9

Commonality Coefficients* for the Regression of Posttest on SES (S), Pretest (P), EdExAG Dummy Codes (D) and Teacher Variables (T) -- Cohort 4, Spring 1973, Reading Total Score Analysis

	SES (S)	Pretest (P)	EdExAG Dummy Codes (D)	Teacher Variables (T)
$U(X_i)$	.00254 (.00335)	.53029 (.69855)	.02448 (.03225)	.00026 (.00034)
C(S,P)	.04814 (.06342)	.04814 (.06342)		
C(S,D)	.00208 (.00274)		.00208 (.00274)	
C(S,T)	.00001 (.00001)			.00001 (.00001)
C(P,D)		.04790 (.06310)	.04790 (.06310)	
C(P,T)		.00857 (.01128)		.00857 (.01128)
C(D,T)			-.00008 (-.00011)	-.00008 (-.00011)
C(S,P,D)	.07561 (.09960)	.07561 (.09960)	.07561 (.09960)	
C(S,P,T)	.00129 (.00170)	.00129 (.00170)		.00129 (.00170)
C(S,D,T)	.00097 (.00128)		.00097 (.00128)	.00097 (.00128)
C(P,D,T)		-.00417 (-.00549)	-.00417 (-.00549)	-.00417 (-.00549)
C(S,P,D,T)	.02124 (.02797)	.02124 (.02797)	.02124 (.02797)	.02124 (.02797)
$R^2(X_i)$	.15188	.72887	.16803	.02808

$R^2(S,P,D,T) = .75912$

* Below each commonality coefficient and in parentheses is the normalized coefficient; that is, the commonality coefficient divided by $R^2(S,P,D,T)$.

TABLE V-10

Commonality Coefficients* for the Regression of Posttest on SES (S), Pretest (P), EdExAG Dummy Codes (D) and Teacher Variables (T) -- Cohort 6, Spring 1972, Reading Total Score Analysis

	SES (S)	Pretest (P)	EdExAG Dummy Codes (D)	Teacher Variables (T)
$U(X_1)$	.00393 (.00524)	.60783 (.81127)	.00648 (.00865)	.00018 (.00024)
C(S,P)	.05701 (.07609)	.05701 (.07609)		
C(S,D)	.00091 (.00122)		.00091 (.00122)	
C(S,T)	.00000 (.00001)			.00000 (.00001)
C(P,D)		.01406 (.01876)	.01406 (.01876)	
C(P,T)		.00537 (.00716)		.00537 (.00716)
C(D,T)			.00068 (.00091)	.00068 (.00091)
C(S,P,D)	.03291 (.04393)	.03291 (.04393)	.03291 (.04393)	
C(S,P,T)	.00095 (.00126)	.00095 (.00126)		.00095 (.00126)
C(S,D,T)	-.00048 (-.00065)		-.00048 (-.00065)	-.00048 (-.00065)
C(P,D,T)		.00083 (.00111)	.00083 (.00111)	.00083 (.00111)
C(S,P,D,T)	.01859 (.02481)	.01859 (.02481)	.01859 (.02481)	.01859 (.02481)
$R^2(X_1)$	.11381	.73754	.07397	.02611

$R^2(S,P,D,T) = .74923$

* Below each commonality coefficient and in parentheses is the normalized coefficient; that is, the commonality coefficient divided by $R^2(S,P,D,T)$.

TABLE V-11

Commonality Coefficients* for the Regression of Posttest on SES (S), Pretest (P), EdExAG Dummy Codes (D) and Teacher Variables (T) -- Cohort 6, Spring 1973, Reading Total Score Analysis

	SES(S)	Pretest (P)	EdExAG Dummy Codes(D)	Teacher Variables(T)
$U(X_1)$	.00143 (.00190)	.60747 (.80538)	.01148 (.01522)	.00033 (.00043)
C(S,P)	.06538 (.08669)	.06538 (.08669)		
C(S,D)	.00002 (.00003)		.00002 (.00003)	
C(S,T)	.00002 (.00003)			.00002 (.00003)
C(P,D)		.00775 (.01027)	.00775 (.01027)	
C(P,T)		.00887 (.01176)		.00887 (.01176)
C(D,T)			.00041 (.00054)	.00041 (.00054)
C(S,P,D)	.03115 (.04130)	.03115 (.04130)	.03115 (.04130)	
C(S,P,T)	.00168 (.00222)	.00168 (.00222)		.00168 (.00222)
C(S,D,T)	.00008 (.00010)		.00008 (.00010)	.00008 (.00010)
C(P,D,T)		-.00005 (-.00006)	-.00005 (-.00006)	-.00005 (-.00006)
C(S,P,D,T)	.01824 (.02419)	.01824 (.02419)	.01824 (.02419)	.01824 (.02419)
$R^2(X_1)$	.11801	.74050	.06909	.02957

$R^2(S,P,D,T) = .75427$

* Below each commonality coefficient and in parentheses is the normalized coefficient; that is, the commonality coefficient divided by $R^2(S,P,D,T)$.

measures of the predictive overlap of that specified combination of variables. This overlap occurred because the predictor/independent variable sets in these data were correlated with each other, an inevitable result in studies of naturally occurring groups of students such as those in EdExAG groups or schools. The analyses reported previously showed that these predictors were confounded or correlated with one another, so the fact that the higher-order commonalities in Tables V-6 through V-11 are nonzero was expected.

Lastly, it should be noted that higher-order commonalities in which the dummy variables are considered indicate among-group confounding. This is because the dummy codes have no within-group variation and because all covariation of these dummy codes with student-level variables will have only among-group components. Conversely, commonalities involving only combinations of student-level predictors (i.e., SES, pretest and the teacher variables) reflect only confounding within groups. (All these technical issues regarding commonality analysis and its properties with multilevel predictor variables are discussed in Appendix IV-E.)

Also presented in Tables V-6 through V-11 are the "normalized" commonality coefficients. Normalized coefficients are computed by dividing each coefficient by the squared multiple correlation based on all predictors [i.e., by $R^2(S,P,D,T)$] and have the property of summing to 1.0. Thus, commonalities from equations with different total R^2s (such as those from different school years or different cohorts) can be compared.

Examination of Tables V-6 through V-11 shows that the overall squared multiple correlations of posttest reading performance with these four sets of variables are quite high. The only exception is for the Spring 1972 analysis for Cohort 1; this finding is undoubtedly due to the fact that the pretest measure for this sample was not a CTBS score (as was the posttest) but was the Mental Age-Language score from the CTMM. Posttest differences among EdExAG groups for Cohorts 1 and 4 were approximately similar, accounting for about 15 percent of the variance in reading performance. As noted previously, the $R^2(X_1)$ for the EdExAG dummy codes provides this measure. EdExAG group differences accounted for a much smaller proportion of the variation in posttest for Cohort 6. One possible explanation for this finding is that, relative to Cohorts 1 and 4, pretest was not highly confounded with EdExAG group membership in the Cohort 6 analysis samples (see Tables V-D-5 and V-D-6 in Appendix V-D). As a result, the second-order commonality coefficients for pretest and the dummy codes [C(P,D)] in the posttest analyses shown in Tables V-6 through V-11 were also smaller for both Cohort 6 data sets. The normalized commonalities show this trend even more clearly.

After group differences with respect to SES, pretest and the teacher variables have been taken into account, the relationship of EdExAG group membership to posttest reading performance drops dramatically, especially for Cohorts 4 and 6 [e.g., for Cohort 4, SP72: $R^2(D) = .16154$, $U(D) = .02024$; for Cohort 6, SP72: $R^2(D) = .07397$, $U(D) = .00648$]. However, substantial differences among EdExAG groups were present for the Cohort 1 samples, even after SES, pretest and teacher differences had been considered. In fact, approximately 7 percent of the variation in the Cohort 1, SP72 posttest reading achievement could be attributed to the unique association of posttest and EdExAG/treatment group membership, while over 13 percent of the SP73 posttest variance was uniquely related to group membership.

Inspection of the normalized first-order coefficients (i.e., uniquenesses) for the two Cohort 1 analyses shows that even though the overall multiple correlations squared were quite different, the unique variance in EdExAG group membership accounted for about the same proportion of the predictable variance in posttest reading achievement ($\approx$ 20 percent) in both analysis data sets, Spring 1972 and Spring 1973. Tables V-6 and V-7 also suggest a possible explanation for the large difference in the overall multiple correlations squared for these two analyses [i.e., $R^2(S,P,D,T) = .33$ versus .66]. The primary reason appears to be the substantially smaller correlation between pretest and posttest in the Cohort 1, SP72 analysis sample ($r = .37$ versus $r = .72$ in the SP73 sample), a result undoubtedly due to the necessity of using the CTMM Mental Age-Language score as a pretest measure in the SP72 analysis. The lower correlation of pretest with SES in the SP72 sample ($r = .17$ versus $r = .41$ for the SP73 sample) along with the smaller pretest/posttest correlation probably also explains why (1) the SES uniqueness for the Cohort 1, SP72 sample was proportionately larger than that for the other five analysis samples; (2) the Cohort 1, SP72 sample's uniqueness for pretest was much smaller than that in the other analysis data sets; and (3) the second-order commonality coefficient for SES and EdExAC group membership in the Cohort 1, SP72 sample [C(S,D)] was much larger than the same coefficient in the other samples; while the third-order coefficient for SES, pretest and group membership [C(S,P,D)] tended to be much smaller.

Although the CTMM Mental Age-Language score is not as reliable as the CTBS Reading Total Expanded Scale Score (see Chapter III, Table III-4), correcting the Cohort 1 pretest/posttest correlations for attenuation on the basis of the formula

$$r_{x_\infty y_\infty} = \frac{r_{xy}}{\sqrt{r_{xx'}}\sqrt{r_{yy'}}} , \qquad \text{(V-2)}$$

where

$r_{x_\infty y_\infty}$ = the pretest (X), posttest (Y) correlation corrected for unreliability in both X and Y,

r_{xy} = the observed correlation between pretest and posttest,

$r_{xx'}$ = the reliability of pretest, and

$r_{yy'}$ = the reliability of posttest.

showed that the differences between the two "corrected" correlations was still sizable (SP72, $r_{x_\infty y_\infty} = .45$; SP73, $r_{x_\infty y_\infty} = .75$). Even though the reliability estimates used in these computations are themselves fallible estimates, the discrepancy between the pretest/posttest correlations in the two Cohort 1 analysis samples is too large to be merely the result of reliability differences with respect to the pretest. Rather, it seems more likely that the CTMM Mental Age-Language and CTBS Reading Total Expanded Scale Score are measuring somewhat different constructs.

Tables V-6 through V-11 also show that the high pretest/posttest correlations tended to increase ever so slightly the older the student; that the pattern of uniquenesses and higher-order commonality coefficients was fairly similar across the SP72 and SP73 analyses for the same cohort; and that, as expected from the SES and pretest analyses reported earlier, the primary sources of predictive overlap in all of the analyses involved confounding of SES and pretest [C(S,P)], pretest and the EdExAG group membership [C(P,D)] and SES, pretest and EdExAG group membership [C(S,P,D)].

One additional issue which examination of Tables V-6 through V-11 brings to mind is: Since previously reported analyses showed that there were substantial differences among EdExAG groups with respect to SES, why is there not a higher degree of confounding of SES with EdExAG group membership [C(S,D)] in these tables? Although a direct explanation of why this occurred is not possible, it seems likely that this result was due to the fact that (1) pretest was highly correlated with posttest and (2) SES was confounded (i.e., correlated) with pretest to about the same degree as it was correlated with posttest. This explanation seems plausible since (1) Tables V-7 through V-11 show that SES is primarily confounded with pretest and with those variable combinations including pretest and (2) SES and EdExAG group membership [C(S,D)] were confounded much more highly in the Cohort 1, SP72 sample (Table V-6) where pretest and posttest are not highly correlated ($r = .37$) and SES is not similarly related to posttest ($r = .41$) and pretest ($r = .17$).

Tables V-6 through V-11 also quantify the degree to which the three teacher variables, Teacher Age, Teacher SES Background and Teaching Qualifications, were related to posttest and the extent to which these variables, taken as a set, were confounded predictors of posttest reading achievement. The results show that these teacher attributes accounted for very little of the variation in posttest performance and that the small associations that were present were almost completely confounded with SES, pretest and EdExAG group membership [C(S,P,D,T)].

Summary of EdExAG group commonality analyses. The results of these commonality analyses indicated that very small differences in reading achievement could be attributed to EdExAG/treatment group differences over and above the effects of SES and initial achievement status (i.e., pretest) for Cohorts 4 and 6, but that there were major group differences in the Cohort 1 analysis samples. For the Cohort 1, SP73 analysis sample, one possible reason for this particular result is that the confounding between SES, pretest and EdExAG group membership [see C(S,P,D) in Tables V-6 through V-11] was smaller in this sample relative to the confounding in the other samples. The variance partitions shown in the commonality tables (V-6 through V-11), however, express only the proportion of criterion variance related to group differences and to the other predictors. Therefore, the correlations between posttest and the key teacher and treatment scales analyzed for this report were examined to see if posttest reading achievement was systematically associated with any particular treatment attributes.

Correlations with posttest. Table V-12 contains these zero-order correlations with posttest. Only the correlations with pretest and SES were both statistically significant ($p < .05$) and consistent (that is, of the same sign) across all of the analysis samples. No other variables showed consistent relationships with posttest across all three cohorts. Most of the correlations that were both statistically significant and of the same sign for the SP72 and SP73 analyses are in the Cohort 6 analysis samples. The Cohort 1 correlations, on the other hand, showed a tendency to be positive in the SP72 analysis sample and negative in the SP73 sample. Most notably, neither the superscales, Degree of Individualization and Level of Innovation, nor the quantity of schooling index, Number of Minutes per Day, had statistically significant and consistent correlations with posttest across all samples.

Unfortunately, these zero-order correlations with posttest did not allow for the possibility that students exposed to different kinds (or levels) of treatments might have been initially different with respect to SES and pretest level and that the correlations were merely reflecting variation which had existed prior to administration of the treatment. One way of making a statistical compensation for these differences is to compute the correlations of the posttest with the part of each variable that could not be predicted from pretest and SES, that is, correlating posttest with an "adjusted" predictor variable. An alternative method involving a different statistical and educational model is to adjust the posttest on the basis of pretest and SES.

TABLE V-12

Correlations of All Key Analysis Variables with Posttest - Reading Total Score Analysis

Level of Correlation	Variable	Cohort 1 SP72	Cohort 1 SP73	Cohort 4 SP72	Cohort 4 SP73	Cohort 6 SP72	Cohort 6 SP73
Student/ Student	Socioeconomic Background	.405	.273	.370	.389	.337	.343
	Pretest	.366	.719	.838	.853	.858	.860
	Teacher Age	-.000	.057	.048	.001	-.105	-.149
	Teacher Socioeconomic Background	-.029	-.064	.071	.163	.077	.070
	Teaching Qualifications	.019	.200	.052	-.009	-.15[illegible]	-.130
EdExAG/ Student	Utilization of Objectives	.131	-.164	.044	-.001	.133	.075
	Individualization in Decision Making	-.052	-.149	-.036	-.040	.078	.027
	Teacher or Locally Developed Materials	-.012	-.078	-.023	-.096	.106	-.078
	Individualization of Instructional Pace	.097	.067	.064	-.080	.089	-.064
	Scheduling Characteristics	.074	-.244	-.066	-.115	.173	.135
	Use of Performance Agreements	.018	-.140	.015	.056	.121	.096
	Classroom Group Organization	-.029	-.022	-.026	-.159	.204	.047
	Teaching Unit Composition	.090	-.131	-.033	-.097	.008	-.040

(continued)

TABLE V-12 (continued)

Level of Correlation	Variable	Cohort 1 SP72	Cohort 1 SP73	Cohort 4 SP72	Cohort 4 SP73	Cohort 6 SP72	Cohort 6 SP73
	Completeness of Instructional Package	.085	-.074	.041	-.052	.115	-.019
	Utilization of Student Evaluation	.089	-.074	.056	-.077	.150	.054
	Degree of Individualization	.058	-.136	.040	-.042	.150	.042
	Level of Innovation	.079	-.193	.009	-.107	.158	.035
	Number of Minutes per Day	-.053	-.088	.041	-.039	-.117	-.210
$r(p < .05)$ - Student/Student Level[1]		.069	.070	.044	.044	.050	.049
$r(p < .05)$ - EdExAG/Student Level[2]		.065	.065	.041	.041	.048	.048

Note: For convenience, all coefficients in this table were truncated to three decimal places--see Appendix V-E for correlations with five-place accuracy.

[1] Absolute value of the coefficient required for a correlation between two student-level variables to be statistically significant ($\alpha \leqq .05$, two-tailed). Due to the large number of degrees of freedom, the critical ratio $z = r/\sigma_r$ where $\sigma_r = 1/\sqrt{N}$ was used.

[2] Absolute value of the coefficient required for a correlation between one student-level variable (Posttest) and an EdExAG-level variable to be statistically significant ($\alpha \leqq .05$, two-tailed)--see equation IV-3 for the computing formula.

It is known that posttest measures of achievement are highly related to initial achievement status as well as to student SES. Treatments and teachers, then, do not influence posttest achievement in an absolute sense. Rather, they are concerned only with achievement that is not expected on the basis of a student's initial achievement level or socioeconomic background. Because commonality analysis and regression analysis involve adjusting the predictor variables on the basis of their relationship with pretest and SES, it was decided that the second alternative method of adjustment would be used to provide a somewhat different analytic perspective. For this reason, the posttest scores in each analysis sample were residualized by subtracting from each posttest score, the posttest score predicted on the basis of that student's pretest and SES level (see equation IV-2 in Chapter IV). This new dependent variable, then, was a measure of reading achievement which quantified how much a student deviated from that level of performance expected of him/her, given his/her particular pretest and SES level. Correlations with this residual variable are shown in Table V-13.

Inspection of these "part correlations" shows that (1) the treatment scales had few statistically significant ($p < .05$) correlations in the SP72 samples for Cohorts 1 and 4; and (2) although quite a few statistically significant correlations ($p < .05$) were present in both Cohort 6 samples, almost all of these correlations tended to be positive in SP72 but negative in SP73. These results, then, indicate that variations in these treatment characteristics are either not linearly related to deviations from expected reading levels or not related in the same way in two consecutive school years.

It should be noted that the change in the sign of the treatment variable correlation with the residual score across SP72 and SP73 analysis samples may be, to an unknown degree, a function of the test score regression-to-the-mean, which is due, at least in part, to the unreliability of the test scores. The tests, however, were fairly reliable in terms of the reliabilities shown in Table III-4 in Chapter III. On the other hand, because many of the same students were in the two analysis data sets for each cohort, many of the same Year 2 test scores were both posttest

TABLE V-13

Correlations of the Key Treatment Variables with Posttest Residualized on the Basis of Pretest and SES - Reading Total Score Analyses

Variable	Cohort 1 SP72	Cohort 1 SP73	Cohort 4 SP72	Cohort 4 SP73	Cohort 6 SP72	Cohort 6 SP73
Utilization of Objectives	.048	-.390	.000	.020	.057	-.095
Individualization in Decision-Making	-.031	-.250	-.049	.053	.048	-.112
Teacher or Locally Developed Materials	.009	.013	-.001	.037	.040	-.108
Individualization of Instructional Pace	.063	-.028	-.007	.019	.101	-.136
Scheduling Characteristics	.120	-.257	-.020	.027	-.011	-.023
Use of Performance Agreements	.070	-.182	-.043	.099	-.050	-.083
Classroom Group Organization	.004	-.038	-.012	.019	-.000	.059
Teaching Unit Composition	.047	-.298	-.028	.001	.118	-.134
Completeness of Instructional Package	.054	-.158	-.036	.007	.070	-.160
Utilization of Student Evaluation	.052	-.270	-.007	.085	.056	-.090
Degree of Individualization	.056	-.293	-.032	.100	.054	-.102
Level of Innovation	.069	-.352	-.027	.063	.057	-.139
Number of Minutes per Day	.010	-.215	.062	.014	.091	-.128
r ($p < .05$) - EdExAG/Student Level[1]	.067	.061	.043	.043	.049	.048

Note: For convenience, all coefficients were truncated to three decimal places.

[1] Absolute value of the coefficient required for a correlation between one student-level variable (posttest residual) and an EdExAG-level variable to be statistically significant ($\alpha \leq .05$, two-tailed)--see equation IV-3 for the basic computing formula. All correlations in this table are mixed-level correlations.

measures in the Year 1/Year 2 analysis and pretest measures in the Year 2/Year 3 analysis. The shift in treatment approaches between Year 2 and Year 3 discussed earlier, then, could have led to the different patterns of achievement suggested by this change in signs.

In spite of all these other explanations for such results, however, the most likely explanation is that treatments were actually not systematically or consistently related to the deviations from reading performance that were expected on the basis of initial reading level and SES across time points and analysis cohorts. These correlational results, then, have not demonstrated that greater innovativeness as defined by these scales or by the two superindices is consistently related to improved student performance in reading.

Analysis of the Two Revised Growth Models

Two multipredictor regression models were developed to examine the hypothesis that growth in reading achievement, as measured by the CTBS Reading Total Expanded Scale Score, was related to Level of Innovation (LI) or Degree of Individualization (DI) as defined by means of the key treatment variables analyzed for this report. In addition to one of these treatment superindices, each model contained, as predictors, a relative measure of the socioeconomic status of each student's home background (SES), a measure of the reading achievement level of each student at the end of the previous school year (pretest), the Number of Minutes per Day that was devoted to language arts instruction and activity (NMIN), and a measure of the qualifications/experience of each child's teacher or teachers (Teaching Qualifications-TQ). Three general analytic procedures were involved in testing these hypotheses. First, the predictive relationships among SES, pretest, Level of Innovation or Degree of Individualization, and Teaching Qualifications were examined via commonality analyses of this four-predictor model. Second, the raw score and standard score regression coefficients were examined and tested for statistical significance. Last, the commonality analysis and regression analysis results were considered in light of the correlations of the predictors with each other and with the posttest reading score.

Growth model commonality analyses. Tables V-14 through V-19 contain the commonality analysis results for the four-set prediction model containing SES (S); pretest (P); the EdExAG measures Level of Innovation (L[illegible]) and Number of Minutes per Day (NMIN); and Teaching Qualifications (T). Normalized coefficients--that is, coefficients divided by the total multiple correlation squared--are also shown in these tables. As noted previously, a general description of how commonality tables such as these may be read has been placed in Appendix V-F.

The overall proportion of posttest variance in reading achievement accounted for by the Level of Innovation model is less than that accounted for by the model that included the EdExAG group dummy codes as treatment-level predictor variables (instead of Level of Innovation and Number of Minutes per Day). It should be remembered, however, that the R^2 for the dummy variable model reflects the maximum proportion of variation in posttest reading performance that could be accounted for by a model containing EdExAG-level predictors instead of dummy codes (see Appendix IV-E for a technical discussion).

One finding that was apparent in Tables V-14 through V-19 was that the difference between the R^2 for the revised growth model and the R^2 for the dummy variable model was related to grade level. Thus, the difference between these two R^2s for the SP73 analysis for Cohort 6 $[R^2(S,P,D,T) - R^2(S,P,E,T) = .75427 - .74937 = .00490]$ was considerably smaller than the same difference (.06572) for the SP72 analysis of Cohort 1. The dummy variable model did differ from this revised model in that two additional teacher variables, Teacher Age and Teacher SES Background, were also present in the dummy variable prediction equation. This is not likely to have produced these differences, however, because all three teacher variables (see Tables V-6 through V-11) accounted for little posttest variance by themselves and very little variance over and above that which was also due to pretest, SES and EdExAG group membership.

One explanation for the difference between the dummy variable model R^2 and the Level of Innovation growth model R^2 is that the dummy variable model indexes all EdExAG group differences, not just those measured by Level of Innovation and Number of Minutes per Day. In addition, simply

TABLE V-14

Commonality Coefficients* for the Level of Innovation Growth Model -- Cohort 1, Spring 1972, Reading Total Score Analysis

Commonality Coefficient	SES(S)	Pretest (P)	EdExAG Variables LI and NMIN(E)**	Teaching Qualifications(T)
$U(X_i)$	.11356 (.43409)	.09510 (.36354)	.00423 (.01616)	.00213 (.00814)
C(S,P)	.04454 (.17027)	.04454 (.17027)		
C(S,E)	.00978 (.03737)		.00978 (.03737)	
C(S,T)	-.00213 (-.00814)			-.00213 (-.00814)
C(P,E)		-.00356 (-.01362)	-.00356 (-.01362)	
C(P,T)		-.00051 (-.00197)		-.00051 (-.00197)
C(E,T)			-.00022 (-.00083)	-.00022 (-.00083)
C(S,P,E)	-.00243 (-.00928)	-.00243 (-.00928)	-.00243 (-.00928)	
C(S,P,T)	.00077 (.00293)	.00077 (.00293)		.00077 (.00293)
C(S,E,T)	.00024 (.00094)		.00024 (.00094)	.00024 (.00094)
C(P,E,T)		.00014 (.00055)	.00014 (.00055)	.00014 (.00055)
C(S,P,E,T)	-.00004 (-.00016)	-.00004 (-.00016)	-.00004 (-.00016)	-.00004 (-.00016)
$R^2(X_i)$	.16429	.13400	.00814	.00038

$R^2(S,P,E,T) = .26159$

$R^2(S,P,D,T)^{+} = .32731$

---

* Below each commonality coefficient and in parentheses is the normalized coefficient; that is, the commonality coefficient divided by $R^2(S,P,E,T)$.

**LI and NMIN are abbreviations for Level of Innovation and Number of Minutes per Day.

+ The multiple correlation squared of posttest with SES, pretest, the dummy variables encoding EdExAG group membership, and the three key teacher variables.

TABLE V-15

Commonality Coefficients* for the Level of Innovation Growth Model -- Cohort 1, Spring 1973, Reading Total Score Analysis

Commonality Coefficient	SES(S)	Pretest (P)	EdExAG Variables LI and NMIN(E)**	Teaching Qualifications(T)
$U(X_i)$	.00076 (.00124)	.43976 (.71608)	.08778 (.14294)	.00563 (.00917)
C(S,P)	.08687 (.14145)	.08687 (.14145)		
C(S,E)	.00015 (.00025)		.00015 (.00025)	
C(S,T)	.00067 (.00109)			.00067 (.00109)
C(P,E)		-.01282 (-.02087)	-.01282 (-.02087)	
C(P,T)		.01432 (.02332)		.01432 (.02332)
C(E,T)			.00306 .00498	.00306 (.00498)
C(S,P,E)	-.02860 (-.04657)	-.02860 (-.04657)	-.02960 (-.04657)	
C(S,P,T)	.01879 (.03060)	.01879 (.03060)		.01879 (.03060)
C(S,E,T)	-.00120 (-.00196)		-.00120 (-.00196)	-.00120 (-.00196)
C(P,E,T)		.00155 (.00252)	.00155 (.00252)	.00155 (.00252)
C(S,P,E,T)	-.00260 (-.00424)	-.00260 (-.00424)	-.00260 (-.00424)	-.00260 (-.00424)
$R^2(X_i)$	.07484	.51727	.04732	.04021

$R^2(S,P,E,T) = .61412$

$R^2(S,P,D,T)^+ = .66135$

---

* Below each commonality coefficient and in parentheses is the normalized coefficient; that is, the commonality coefficient divided by $R^2(S,P,E,T)$.

**LI and NMIN are abbreviations for Level of Innovation and Number of Minutes per Day.

+ The multiple correlation squared of posttest with SES, pretest, the dummy variables encoding EdExAG group membership, and the three key teacher variables.

TABLE V-16

Commonality Coefficients* for the Level of Innovation Growth Model -- Cohort 4, Spring 1972, Reading Total Score Analysis

Commonality Coefficient	SES(S)	Pretest (P)	EdExAG Variables LI and NMIN(E)**	Teaching Qualifications(T)
$U(X_i)$	.00801 (.01123)	.56461 (.79224)	.00150 (.00211)	.00005 (.00008)
C(S,P)	.13514 (.18962)	.13514 (.18962)		
C(S,E)	.00008 (.00011)		.00008 (.00011)	
C(S,T)	.00015 (.00021)			.00015 (.00021)
C(P,E)		.00318 (.00446)	.00318 (.00446)	
C(P,T)		.00517 (.00726)		.00517 (.00726)
C(E,T)			.00002 (.00002)	.00002 (.00002)
C(S,P,E)	-.00264 (-.00371)	-.00264 (-.00371)	-.00264 (-.00371)	
C(S,P,T)	-.00216 (-.00304)	-.00216 (-.00304)		-.00216 (-.00304)
C(S,E,T)	.00008 (.00011)		.00008 (.00011)	.00008 (.00011)
C(P,E,T)		.00052 (.00074)	.00052 (.00074)	.00052 (.00074)
C(S,P,E,T)	-.00102 (-.00143)	-.00102 (-.00143)	-.00102 (-.00143)	-.00102 (-.00143)
$R^2(X_i)$	.13762	.70279	.00171	.00280

R^2(S,P,E,T) = .71267

R^2(S,P,D,T)+ = .73216

---

* Below each commonality coefficient and in parentheses is the normalized coefficient; that is, the commonality coefficient divided by R^2(S,P,E,T).

**LI and NMIN are abbreviations for Level of Innovation and Number of Minutes per Day.

+ The multiple correlation squared of posttest with SES, pretest, the dummy variables encoding EdExAG group membership, and the three key teacher variables.

TABLE V-17

Commonality Coefficients* *for the Level of Innovation Growth Model--* Cohort 4, Spring 1973, Reading Total Score Analysis

Commonality Coefficient	SES(S)	Pretest (P)	EdExAG Variables LI and NMIN(E)**	Teaching Qualifications(T)
$U(X_i)$	.00588 (.00799)	.57804 (.78556)	.00135 (.00184)	.00000 (.00000)
C(S,P)	.13580 (.18455)	.13580 (.18455)		
C(S,E)	-.00035 (-.00047)		-.00035 (-.00047)	
C(S,T)	.00003 (.00004)			.00003 (.00004)
C(P,E)		.00429 (.00584)	.00429 (.00584)	
C(P,T)		.00006 (.00008)		.00006 (.00008)
C(E,T)			.00001 (.00001)	.00001 (.00001)
C(S,P,E)	.01072 (.01456)	.01072 (.01456)	.01072 (.01456)	
C(S,P,T)	.00036 (.00049)	.00036 (.00049)		.00036 (.00049)
C(S,E,T)	.00004 (.00005)		.00004 (.00005)	.00004 (.00005)
C(P,E,T)		.00019 (.00026)	.00019 (.00026)	.00019 (.00026)
C(S,P,E,T)	-.00059 (-.00080)	-.00059 (-.00080)	-.00059 (-.00080)	-.00059 (-.00080)
$R^2(X_i)$	.15188	.72887	.01566	.00009

$R^2(S,P,E,T) = .73583$

$R^2(S,P,D,T)^+ = .75912$

---

* Below each commonality coefficient and in parentheses is the normalized coefficient; that is, the commonality coefficient divided by $R^2(S,P,E,T)$.

**LI and NMIN are abbreviations for Level of Innovation and Number of Minutes per Day.

+ The multiple correlation squared of posttest with SES, pretest, the dummy variables encoding EdExAG group membership, and the three key teacher variables.

TABLE V-18

Commonality Coefficients* for the Level of Innovation Growth Model--Cohort 6, Spring 1972, Reading Total Score Analysis

Commonality Coefficient	SES(S)	Pretest (P)	EdExAG Variables LI and NMIN(E)**	Teaching Qualifications(T)
$U(X_1)$	.00467 (.00628)	.61576 (.82675)	.00288 (.00386)	.00015 (.00020)
C(S,P)	.07078 (.09503)	.07078 (.09503)		
C(S,E)	-.00036 (-.00048)		-.00036 (-.00048)	
C(S,T)	.00019 (.00026)			.00019 (.00026)
C(P,E)		.00613 (.00823)	.00613 (.00823)	
C(P,T)		.00356 (.00478)		.00356 (.00478)
C(E,T)			-.00012 (-.00017)	-.00012 (-.00017)
C(S,P,E)	.01997 (.02681)	.01997 (.02681)	.01997 (.02681)	
C(S,P,T)	.00450 (.00604)	.00450 (.00604)		.00450 (.00604)
C(S,E,T)	-.00015 (-.00020)		-.00015 (-.00020)	-.00015 (-.00020)
C(P,E,T)		.00263 (.00353)	.00263 (.00353)	.00263 (.00353)
C(S,P,E,T)	.01421 (.01909)	.01421 (.01909)	.01421 (.01909)	.01421 (.01909)
$R^2(X_1)$	.11381	.73754	.04519	02497

$R^2(S,P,E,T) = .74480$

$R^2(S,P,D,T)^+ = .74923$

---

* Below each commonality coefficient and in parentheses is the normalized coefficient; that is, the commonality coefficient divided by $R^2(S,P,E,T)$.

**LI and NMIN are abbreviations for Level of Innovation and Number of Minutes per day.

+ The multiple correlation squared of posttest with SES, pretest, the dummy variables encoding EdExAG group membership, and the three key teacher variables.

TABLE V-19

Commonality Coefficients* for the Level of Innovation Growth Model -- Cohort 6, Spring 1973, Reading Total Score Analysis

Commonality Coefficient	SES(S)	Pretest (P)	EdExAG Variables LI and NMIN(E)**	Teaching Qualifi-cations(T)
$U(X_i)$	.00120 (.00160)	.62065 (.82823)	.00729 (.00972)	.00022 (.00030)
C(S,P)	.07636 (.10190)	.07636 (.10190)		
C(S,E)	.00038 (.00051)		.00038 (.00051)	
C(S,T)	-.00010 (-.00014)			-.00010 (-.00014)
C(P,E)		.00080 (.00106)	.00080 (.00106)	
C(P,T)		.00112 (.00150)		.00112 (.00150)
C(E,T)			-.00019 (-.00026)	-.00019 (-.00026)
C(S,P,E)	.02573 (.03434)	.02573 (.03434)	.02573 (.03434)	
C(S,P,T)	.00467 (.00623)	.00467 (.00623)		.00467 (.00623)
C(S,E,T)	.00008 (.00010)		.00008 (.00010)	.00008 (.00010)
C(P,E,T)		.00147 (.00197)	.00147 (.00197)	.00147 (.00197)
C(S,P,E,T)	.00969 (.01293)	.00969 (.01293)	.00969 (.01293)	.00969 (.01293)
$R^2(X_i)$	.11801	.74050	.04525	.01696

$R^2(S,P,E,T) = .74937$

$R^2(S,P,D,T)^{+} = .75427$

---

* Below each commonality coefficient and in parentheses is the normalized coefficient; that is, the commonality coefficient divided by $R^2(S,P,E,T)$.

**LI and NMIN are abbreviations for Level of Innovation and Number of Minutes per Day.

+ The multiple correlation squared of posttest with SES, pretest, the dummy variables encoding EdExAG group membership, and the three key teacher variables.

including quantitative treatment variables as predictors in the growth model permitted only linear associations among the student-level variables (posttest, pretest and SES) and the treatment measures. The commonality analysis results for the dummy variable model were therefore expected to differ from the growth model results because including the group membership dummy codes as a predictor set results in the computation of all among-group overlap, not just that which is linear. Inspection of Tables V-14 through V-19, for example, does show that in relation to the dummy variable model the relative degree of confounding of SES, pretest, and SES plus pretest with the treatment variables [C(S,E), C(P,E), C(S,P,E)] is less in the revised growth models than in the dummy variable model [i.e., C(S,D), C(P,D), C(S,P,D)].

The proportion of criterion variance attributable to the two treatment variables, Level of Innovation and Number of Minutes per Day, was notable only for the Cohort 1, SP73 analysis [$R^2(E) = .04732$] and the SP72 [$R^2(E) = .04519$] and SP73 [$R^2(E) = .04525$] analyses for Cohort 6. Only in the SP73 analysis for Cohort 1 was the uniqueness for this two-variable predictor set worthy of note [U(E) = .08778].

The fact that in the Cohort 1, SP73 analysis sample the uniqueness for the treatment variables was almost twice the multiple correlation squared of those same variables alone with the criterion suggests that some suppressor-like effects might be present in these data. The intercorrelations among these variables (see Table V-20) show that both Level of Innovation and Number of Minutes per Day have negative correlations with posttest and with each other but primarily positive correlations with the other predictors. Although interpretations of effects such as these are somewhat risky, one explanation for this finding is that students who are in treatments characterized by a higher than average Level of Innovation and Number of Minutes per Day (higher relative to all of the students in this analysis sample) tend to be of higher than average SES and to have greater than average pretest scores; nevertheless, they end up with lower than average reading scores at posttest time. The predominance of negative correlations between posttest (or posttest residualized on the basis of pretest and SES) and most of the individual treatment EdExAG scales

TABLE V-20

Intercorrelations among the Level of Innovation Model Variables for the Cohort 1, Spring 1973 Analysis Sample

Variable	Variable Abbreviation					
	Posttest	SES	Pretest	LI	NMIN	TQ
Posttest	-	.274	.719	-.194	-.088	.201
SES	.274	-	.405	.139	.175	.168
Pretest	.719	.405	-	.074	.090	.155
Level of Innovation	-.194	.139	.074	-	-.054	.019
Number of Minutes per Day	-.088	.175	.090	-.054	-	-.063
Teaching Qualifications	.201	.168	.155	.019	-.063	-

also bolsters the argument that a greater emphasis on innovation as measured here by the Level of Innovation superindex is related to depressed reading performance for students in the Cohort 1, SP73 analysis sample. Although other analysis samples do not demonstrate this same pattern of results, it is quite provocative that this trend was observed in the only analysis sample in which educationally substantial treatment effects were present.

Tables V-G-1 through V-G-6 in Appendix V-G show the results of the commonality analyses for the Degree of Individualization growth model. The high correlations (ranging from .89 to .98) between Degree of Individualization and Level of Innovation (see Appendix V-E) in these samples and the fact that the two superindices tended to have somewhat similar correlations with the other predictors in the model resulted in a great amount of similarity with respect to the commonality analysis results. The high correlation between the two superindices was expected, of course, since four of the EdExAG scale scores forming Level of Innovation were used to define Degree of Individualization. Because the tables in Appendix V-G are so similar to Tables V-14 through V-19, which depict the results for the Level of Innovation model, the reader is referred to this earlier discussion.

Growth model regression analyses. The regression coefficients for the regression of posttest on the five predictor variables in the Level of Innovation growth model are found in Table V-21. Coefficients for the Degree of Individualization model are shown in Table V-22. As discussed in Chapter IV, it can be shown derivationally that the regression coefficient for an independent variable is the influence of that variable (controlled for differences with respect to initial achievement status) on growth when initial achievement status has also been included as a predictor variable in the model (Werts and Linn, 1970). Assuming that this holds true for more than one additional predictor of achievement, then, the regression coefficients in these tables may be viewed as indices of the extent to which growth in reading achievement is related to these other indices, each variable being statistically controlled for the other variables in the model.

Inspection of the coefficients in Table V-21 and Table V-22 shows that, as expected, the regression coefficient for SES was statistically significant in all analysis samples except that for Cohort 1, SP73. The commonality analyses discussed previously, however, showed that the part correlation squared of SES [i.e., the uniqueness for SES, U(SES)] with posttest was extremely small in all analysis samples except that for Cohort 1, SP72. And, as explained, this exception was probably due to the fact that the correlation of SES with pretest (from the CTMM for the Cohort 1, SP72 analysis) was smaller than it was for other analysis samples and that the predictability of posttest due to SES was, therefore, less highly confounded with pretest. The confounding of SES with oth[illegible] variables in the prediction model [most notably, C(S,P) and C(S,P,E) in the commonality tables] suggests that, in addition to its "direct" association with growth reflected by its regression coefficient, SES may also ha[illegible]ad some "indirect" effects. The nature of these indirect influences is, of course, unknown and is primarily a matter of conjecture. One of the most obvious speculations is that SES probably was additionally related to posttest achievement because of its influence on pretest.

TABLE V-21

Regression Coefficients for the Level of Innovation Growth Model - Reading Total Score Analyses[1]

Cohort	Posttest Year	Predictor	r	B	β	F
1	SP72	SES	.41	2.23	.35	120.72*
		Pretest	.37	1.89	.32	101.10*
		Level of Innovation	.08	.97	.07	4.38*
		NMIN	-.05	.04	.02	.38
		TQ	.02	-.38	-.05	2.27
		(Constant)		-47.10		(df = 1; 785)
1	SP73	SES	.27	.21	.03	1.50
		Pretest	.72	.76	.73	867.26*
		Level of Innovation	-.19	-5.33	-.26	132.42*
		NMIN	-.09	-.32	-.17	53.31*
		TQ	.20	.80	.08	11.10*
		(Constant)		185.86		(df = 1; 761)
4	SP72	SES	.37	.68	.10	53.22*
		Pretest	.84	.86	.81	3823.95*
		Level of Innovation	.01	-.33	-.02	2.81
		NMIN	.04	.09	.04	8.45*
		TQ	.05	-.07	-.01	.37
		(Constant)		29.22		(df = 1; 1947)
4	SP73	SES	.39	.64	.08	42.73*
		Pretest	.85	.94	.83	4199.00*
		Level of Innovation	-.11	.81	.04	9.40*
		NMIN	-.04	.05	.02	1.84
		TQ	-.01	.01	.00	.00
		(Constant)		-23.62		(df = 1; 1919)

(continued)

TABLE V-21 (continued)

Cohort	Posttest Year	Predictor	r	B	β	F
6	SP72	SES	.34	.60	.08	27.73*
		Pretest	.86	.90	.84	3653.00*
		Level of Innovation	.16	.32	.02	2.02
		NMIN	-.12	.37	.05	11.64*
		TQ	-.16	-.13	-.01	.89
		(Constant)		-1.68		(df = 1; 1514)
6	SP73	SES	.34	.35	.04	7.38*
		Pretest	.86	.98	.85	3828.45*
		Level of Innovation	.04	-1.25	-.07	25.35*
		NMIN	-.21	-.57	-.06	15.77*
		TQ	-.13	.20	.02	1.37
		(Constant)		29.08		(df = 1; 1546)

[1] Notation and abbreviation:

r = zero-order correlation with posttest
B = raw score regression coefficient
β = standard score regression coefficient
F = conventional F test of the regression coefficient
NMIN = EdExAG scale, Number of Minutes per Day (spent on language arts)
TQ = teacher scale, Teaching Qualifications

$^*p \leqq .05$

TABLE V-22

Regression Coefficients for the Degree of Individualization Growth Model - Reading Total Score Analyses[1]

Cohort	Posttest Year	Predictor	r	B	β	F
1	SP72	SES	.41	2.24	.36	123.54*
		Pretest	.37	1.88	.31	100.06*
		Degree of Individualization	.06	1.98	.06	3.22
		NMIN	-.05	.02	.01	.12
		TQ	.02	-.42	-.05	2.72
		(Constant)		-38.12		(df = 1; 785)
1	SP73	SES	.27	.17	.03	.96
		Pretest	.72	.76	.74	852.32*
		Degree of Individualization	-.14	-10.69	-.24	102.01*
		NMIN	-.09	-.35	-.18	60.68*
		TQ	.20	.80	.08	10.58*
		(Constant)		166.18		(df = 1; 761)
4	SP72	SES	.37	.67	.10	54.10*
		Pretest	.84	.86	.81	3845.08*
		Degree of Individualization	.04	-.78	-.02	2.75
		NMIN	.04	.09	.03	8.05*
		TQ	.05	-.05	-.01	.19
		(Constant)		26.92		(df = 1; 1946)
4	SP73	SES	.39	.64	.08	42.99*
		Pretest	.85	.94	.83	4274.28*
		Degree of Individualization	-.04	2.67	.05	20.96*
		NMIN	-.04	.04	.01	1.45
		TQ	-.01	.03	.00	.05
		(Constant)		-29.82		(df = 1; 1919)

(continued)

TABLE V-22 (continued)

Cohort	Posttest Year	Predictor	r	B	β	F
6	SP72	SES	.34	.60	.08	28.38*
		Pretest	.86	.90	.84	3652.53*
		Degree of Individualization	.15	.71	.02	1.51
		NMIN	-.12	.37	.05	11.79*
		TQ	-.16	-.14	-.01	.96
		(Constant)		-.96		(df = 1; 1514)
6	SP73	SES	.34	.40	.05	8.96*
		Pretest	.86	.97	.85	3829.85*
		Degree of Individualization	.04	-3.10	-.07	27.12*
		NMIN	-.21	-.50	-.05	11.61*
		TQ	-.13	.08	.01	.21
		(Constant)		31.71		(df = 1; 1546)

[1] Notation and abbreviations:

r = zero-order correlation with posttest
B = raw score regression coefficient
β = standard score regression coefficient
F = conventional F test of the regression coefficient
NMIN = EdExAG scale, Number of Minutes per Day (spent on language arts)
TQ = teacher scale, Teaching Qualifications

*$p \leq .05$

A number of the regression coefficients for Level of Innovation (Table V-21) and Degree of Individualization (Table V-22) were statistically significant ($p < .05$), suggesting that these characteristics, corrected for all of the other variables in the models, were related to growth in reading achievement for the populations of which each sample is representative. The apparent impact of these variables, however, is not consistent across analysis samples involving the same cohort or across cohorts. As with the analyses reported earlier, the "direct" impact of either of these two EdExAG/treatment superindices is negative in the only data set where the unique association of these scales with posttest was substantial: Cohort 1, Spring 1973. And, in the only Cohort 6 analysis sample (SP73) in which there was both a statistically significant regression coefficient and a nontrivial proportion of posttest variance related to Level of Innovation and Number of Minutes per Day [Table V-19, $R^2(E) = .05$] or to Degree of Individualization and Number of Minutes per Day [Appendix V-G, Table V-G-6, $R^2(E) = .05$], the regression coefficients were negative. The regression coefficients for Number of Minutes per Day spent in language arts class activities were also negative and statistically significant ($p < .05$) in the Cohort 1, SP73 and Cohort 6, SP73 analysis samples.

Growth model summary. The irregularity of the overall pattern of regression coefficients and the small uniquenesses for Level of Innovation plus Number of Minutes per Day in all samples but one suggest that no consistent overall relation existed between innovative intensity (as indexed by Level of Innovation) and posttest reading performance or between individualization emphasis (as indexed by Degree of Individualization) and posttest reading performance. Teaching Qualifications was not related to posttest performance to any meaningful degree and received a statistically significant regression coefficient in only one analysis sample (the coefficient was positive). The part correlations in which posttest was adjusted on the basis of pretest and SES (i.e., the posttest residual) also did not suggest that consistently positive and substantial associations of innovative intensity with posttest reading achievement were present. Since the same general lack of meaningful impact was observed whether or not the criterion or the predictors were residualized, the hypothesis that a higher Level of Innovation is positively and consistently related to student reading achievement is not supported by these data.

Examination of Unusually Effective EdExAG Groups

As suggested in Chapter IV, it is possible that overall analyses based on the total number of students in each cohort could have failed to identify important correlates of reading achievement by not being sensitive to treatment characteristics that were dramatically influential for only a small subset of students in an analysis sample. Two analytic approaches were used to examine this possibility. The first method consisted of examining the attributes of the EdExAG groups whose students, on the average, did much better or much worse on the posttest than was predicted from their pretest level and SES. The second approach involved comparing and contrasting student characteristics with the treatment and teacher attributes of individual students who exhibited unusual growth or lack of growth in two consecutive school years.

This section will describe the results of the outlier group analyses, while the discussion of the correlates of unusual student growth patterns will be presented in the following section. It should be noted that both of these procedures were exploratory in nature. By design, they involved a small and highly select set of EdExAG groups or individual students; hence inferential statistics were deliberately omitted from any of the comparisons.

For purposes of the outlier group analyses, the EdExAG groups with the greatest positive and the greatest negative mean residual scores in each year were selected for examination, after eliminating those groups with very few students. Two independent analyses were performed for each cohort, the Spring 1972 posttest analysis and the Spring 1973 posttest analysis. For illustration purposes, a table showing the key analysis variables is presented in this section for each analysis year for Cohort 1. The outlier group analysis tables for Cohorts 4 and 6 can be found in Appendix V-H.

To facilitate comparing the groups, the greatest negative outlier group (the poorest performing EdExAG) is presented immediately adjacent to the greatest positive outlier group. As discussed previously, each cohort is numbered in terms of the grade that the students were in during Year 1 of the study, the 1970-71 school year. For example, the students in Cohort 1 were in grade 2 at the time of the Spring 1972 posttest analysis and in grade 3 for the Spring 1973 analysis. Also, it should be mentioned at the outset that outlier groups were identified relative to the mean posttest residual scores of all the other EdExAG groups in a particular analysis sample. It is possible, therefore, for a negative outlier group in one analysis sample to show a yearly gain that is quite respectable relative to overall national norms. On the other hand, a positive outlier group in a different analysis sample might show a rather mediocre gain in comparison to national norms.

Four positive outlier groups and two negative outlier groups were identified for the Cohort 1, Spring 1972 Reading Total analysis, while three positive and three negative outlier groups were selected for the Spring 1973 analysis. Tables V-23 and V-24 present the key analysis variables discussed in the present report. All positive outliers in the Spring 1972 analysis had mean pretest and posttest scores above national norms, whereas the two negative outliers were at or above the national norm on pretest but below the norm on posttest. The mean difference between the groups on posttest is indicated roughly by the fact that the average second grade student in the high performing EdExAGs obtained a posttest score corresponding to a grade equivalent ranging between 3.0 and 3.4 (depending on the group); mean posttest scores in the two low performing EdExAGs ranged between 1.4 and 2.4 grade equivalents. For the 1973 analysis two of the three positive outliers were above national norms on pretest and all three were above on posttest, while two of the three negative outliers were below national norms on both pretest and posttest. A gross indicator of the difference in yearly growth between the performance of students in each group is shown by the grade equivalent scores for the "average" student in each group--the yearly gains of the positive outlier groups ranged between 1.7 and 3.2 grade equivalents during 1973

TABLE V-23

Comparison of Key Analysis Variables for Positive and Negative Outlier EdExAG Groups-- Cohort 1, Spring 1972, Reading Total Score Analysis

Variable		Greatest Negative Outliers		Greatest Positive Outliers			
Number of Students		19	14	73	43	32	101
Posttest Residualized for Pretest and SES	$\bar{X}$	-26.16	-52.22	22.29	18.38	11.42	10.58
	SD	62.50	52.62	39.48	46.35	50.00	46.47
	g[1]	-1.19	-0.65	-0.48	-0.03	-0.83	-0.33
Pretest (National Norm MA = 85)	$\bar{X}$	95.71	84.66	94.72	96.76	93.78	92.53
	SD	10.07	11.94	9.50	8.45	10.93	8.54
	g	-0.01	-0.03	-1.42	-0.23	-0.88	-0.79
Posttest (Interpolated National Norm = 321.5)	$\bar{X}$	309.30	251.49	363.20	360.22	337.31	357.45
	SD	60.58	40.58	44.36	53.00	52.34	54.17
	g	-0.01	-0.35	-0.18	0.17	-0.33	-0.23
SES	$\bar{X}$	96.48	91.34	99.73	98.47	93.78	104.21
	SD	7.81	5.86	9.25	6.50	8.73	9.01
	g	1.06	-1.16	-0.02	0.76	0.02	-0.43
Ethnic Group[2] (white=1; non-white=0)	$\bar{X}$	0.47	0.64	0.99	0.81	0.56	0.87
Sex[2] (girl=1; boy=0)	$\bar{X}$	0.58	0.64	0.56	0.44	0.56	0.59
Utilization of Objectives		3.00	2.33	3.00	3.00	3.00	3.00
Individualization in Decision Making		1.00	1.00	1.00	1.00	1.00	1.67

(continued)

[1] Index of skewness

$$g = \frac{\frac{\Sigma x^3}{N}}{\frac{\Sigma x^2}{N}\sqrt{\frac{\Sigma x^2}{N}}}$$

, where g = skewness index,

x = deviation of a score from $\bar{X}$, where $\bar{X}$ is an EdExAG group mean for variable X, and

N = number of students.

[2] For binary coded variables the $\bar{X}$ is equal to the proportion of students coded with a "1."

TABLE V-23 (continued)

Variable	Greatest Negative Outliers		Greatest Positive Outliers			
Teacher or Locally Developed Materials	1.00	2.00	1.50	1.00	1.50	2.00
Individualization of Instructional Pace	2.00	1.00	3.00	2.00	3.00	2.00
Scheduling Characteristics	1.00	1.00	3.00	1.00	3.00	2.00
Use of Performance Agreements	1.00	1.00	2.00	1.00	3.00	2.00
Classroom Group Organization	2.00	2.00	2.00	3.00	3.00	2.00
Teaching Unit Composition	2.50	2.00	1.50	2.00	2.00	3.00
Completeness of Instructional Package	2.00	1.50	3.00	1.00	2.50	2.50
Utilization of Student Evaluation	2.00	2.67	2.00	2.33	3.00	3.00
Level of Innovation	17.50	16.50	22.00	17.33	25.00	23.17
Degree of Individualization	6.00	5.67	8.00	6.33	10.00	8.67
Number of School Days per Year	180.00	180.00	180.00	180.00	180.00	200.00
Number of Minutes per Day	90.00	105.00	90.00	120.00	70.00	90.00
Number of Teachers $\bar{X}$	4.00	1.00	1.00	2.00	1.00	6.00
Teacher Age $\bar{X}$	29.50	24.00	25.32	35.00	46.88	29.67
Teacher Socioeconomic Background $\bar{X}$	110.35	117.86	108.32	104.78	95.05	95.21
Teaching Qualifications $\bar{X}$	89.91	86.49	87.74	100.29	108.05	98.76

TABLE V-24

Comparison of Key Analysis Variables for Positive and Negative Outlier EdExAG Groups-- Cohort 1, Spring 1973, Reading Total Score Analysis

Variable		Greatest Negative Outliers			Greatest Positive Outliers		
Number of Students		62	17	112	41	36	40
Posttest Residualized for Pretest and SES	$\bar{X}$	-11.08	-16.78	-32.95	66.86	40.66	14.46
	SD	39.24	22.23	36.60	52.87	52.00	37.71
	g[1]	-0.01	1.01	-0.25	0.14	0.06	-0.26
Pretest (Interpolated National Norm = 321.5)	$\bar{X}$	305.72	300.79	357.89	336.28	304.67	365.45
	SD	58.21	26.13	52.72	58.65	62.70	60.58
	g	0.78	-1.06	-0.22	-0.74	-0.17	-0.08
Posttest (Interpolated National Norm = 375.5)	$\bar{X}$	365.90	356.26	381.91	466.93	416.65	434.28
	SD	59.02	18.90	59.98	76.39	72.21	57.90
	g	0.32	0.30	-0.06	-0.14	0.49	-0.08
Reading Interest SP73	$\bar{X}$	101.40	96.94	99.34	106.09		100.79
	SD	9.57	10.97	10.39	7.14		11.38
	g	-0.53	-0.38	-0.01	-0.38		-0.61
SES	$\bar{X}$	93.26	94.82	103.85	93.10	94.63	109.18
	SD	4.65	7.78	9.23	5.18	7.50	5.68
	g	1.56	-0.41	-0.43	1.04	0.38	-0.53
Ethnic Group[2] (white=1; nonwhite=0)	$\bar{X}$	0.95	0.82	0.86	0.95	0.45	0.98
Sex[2] (girl=1;boy=0)	$\bar{X}$	0.52	0.53	0.62	0.39	0.70	0.40

(continued)

Note: Blank entries indicate no data collected.

[1] Index of skewness

$$g = \frac{\frac{\Sigma x^3}{N}}{\frac{\Sigma x^2}{N}\sqrt{\frac{\Sigma x^2}{N}}}$$

, where g = skewness index,

x = deviation of a score from $\bar{X}$, where $\bar{X}$ is an EdExAG group mean for variable X, and

N = number of students.

[2] For binary coded variables the $\bar{X}$ is equal to the proportion of students coded with a "1."

TABLE V-24 (continued)

Variable	Greatest Negative Outliers			Greatest Positive Outliers		
Utilization of Objectives	3.00	3.00	3.00	1.00	1.00	3.00
Individualization in Decision Making	2.00	1.00	1.67	1.00	1.00	1.33
Teacher or Locally Developed Materials	2.00	1.00	2.00	2.00	2.00	1.00
Individualization of Instructional Pace	2.00	1.00	2.00	2.00	2.00	2.00
Scheduling Characteristics	3.00	2.00	2.00	1.00	1.00	2.00
Use of Performance Agreements	3.00	1.00	2.00	1.00	1.00	1.00
Classroom Group Organization	3.00	2.00	2.00	2.00	2.00	3.00
Teaching Unit Composition	2.50	1.00	3.00	1.00	1.00	2.00
Completeness of Instructional Package	2.00	1.50	2.00	1.50	1.50	1.00
Utilization of Student Evaluation	2.67	2.67	3.00	2.33	1.00	2.33
Level of Innovation	25.17	16.17	22.67	14.83	13.50	18.67
Degree of Individualization	9.67	5.67	8.67	6.33	5.00	6.66
Number of School Days per Year	180.00	180.00	200.00	180.00	180.00	180.00
Number of Minutes per Day	40.00	120.00	150.00	90.00	50.00	60.00
Teacher/Student Contact Ratio[3]	3.30	3.85	6.67	3.45	3.87	3.33

(continued)

[3] Relative measure of teacher/student contact equal to the percentage of teacher time per individual student.

TABLE V-24 (continued)

Variable		Greatest Negative Outliers			Greatest Positive Outliers		
Number of Teachers	$\bar{X}$	1.00	1.88	1.00	1.00	1.00	1.60
Teacher Age	$\bar{X}$	44.16	28.76	35.18	30.00	60.00	44.00
Teacher Socioeconomic Background	$\bar{X}$	100.34	99.92	96.93	91.26	108.04	90.74
Teaching Qualifications	$\bar{X}$	97.18	90.22	101.22	100.29	109.96	105.26

as compared with gains of from .5 to 1.1 grade equivalents for the negative outliers.

Considering student characteristics, the high performing groups in 1972, in general, were higher in SES and contained a higher proportion of white students than did the low performing groups. In 1973, the SES level of the positive and negative outlier groups was about the same. Two of the three positive outlier groups had a higher proportion of whites and a lower proportion of girls than did the greatest negative outlier; however, the third high performing group had a higher proportion of girls and a majority of nonwhite students. In general, the average Reading Interest of students was slightly higher in the positive outlier groups in 1973, the first year such information was obtained for Cohort 1.

As can be seen in Tables V-23 and V-24, there is no compelling evidence of treatment characteristics that consistently differentiated the positive outliers from the negative outliers for both years. Considering our two superscales, Level of Innovation and Degree of Individualization, we find that the positive outlier groups typically obtained higher scores than did the low performing groups in 1972. However, in 1973 the situation is reversed--two of the three high performing groups displayed a lower Level of Innovation and a lower Degree of Individualization than did the low performing EdExAG groups. Although the differences are slight, these general trends are also apparent with respect to the individual scales. The only discernible difference between the negative and positive outliers that applied to both years was the tendency for most of the high performing groups to have had teachers who scored higher on the Teaching Qualifications scale and lower on the Teacher SES Background scale than did the teachers of the low performing EdExAG groups.

The other profile variables listed in Appendix IV-H were also inspected to determine if there were any systematic differences between the positive outlier groups and the negative outliers. However, although a number of variables appeared to differentiate between the two groups in a single year, there were none that clearly and consistently differentiated in both years.

For Cohort 4, six outlier groups, three positive and three negative, were identified for each of the two years. Tables V-H-1 and V-H-2 in Appendix V-H contain descriptive statistics on these groups with respect to each of the key variables analyzed for this report.

Of the six EdExAGs included in the Spring 1972 analysis, only one negative outlier had a mean pretest score below the national norm, while the other two negative outlier groups and all three positive outlier groups scored above the norm on pretest. On the posttest, all three negative outliers were below the norm, and all three positive outliers were well above it. A rough indication of the difference between the positive and the negative outliers in 1972 is shown by the average yearly gains of students in these two groups. The average student in the positive outlier groups gained between 1.3 and 1.5 grade equivalents; those in the low performing EdExAG groups gained between .2 and .7 grade equivalents. For the 1973 analysis, all but one of the outlier groups had a mean pretest well above the national norm; the exception was a positive outlier whose mean was rather close to the norm. On the Spring 1973 posttest, all positive outliers and one negative outlier were above the norm. The gain of the average student in grade equivalent units was again calculated for purposes of illustrating the differences in student growth between the positive and negative outlier groups. Students in the high performing EdExAG groups averaged a gain of about 1.9 grade equivalents; the average gain for students in the low performing EdExAG groups ranged from 0 to .3 grade equivalents.

Looking at student characteristics for the Spring 1972 analysis, we found the positive outlier groups were more often higher in mean SES than the negative outliers. In terms of ethnic distribution, one of the negative outliers had only 75 percent white students; the other negative outliers and all three positive outliers were over 90 percent white. Both positive and negative outlying groups tended to have slightly more boys than girls. In terms of Reading Interest, there were generally only slight differences between the high performing and the low performing groups in Spring 1971 and in Spring 1972. The greatest change across school years was a decrease in Reading Interest occurring in one of the negative outliers.

In the Spring 1973 analysis the SES differences were rather mixed and the groups were generally similar in sex distribution. However, the positive outliers all had high proportions of white students (.89 to .97), whereas the negative outlier groups ranged widely (.40 to 1.00).

As was true with the outlier groups from Cohort 1, there appears to be no evidence in Cohort 4 of treatment characteristics that consistently differentiate the high performing groups from the low performing groups for both years. Although the positive outliers in 1973 are characterized generally by a higher Level of Innovation and a greater Degree of Individualization than the negative outlier groups, no clear differences on these superscales were present in 1972. The same trends were not as apparent among the individual scales making up these two superscales. The most notable of the treatment differences in the 1972 analyses was the greater amount of instruction time spent on language arts by two of the three positive outliers. This trend, however, was not present in the 1973 analyses. In terms of teacher characteristics, the higher performing groups in 1973 typically had teachers who were older, but the results for 1972 were mixed.

Looking at the other profile variables listed in Appendix IV-H, again—as in Cohort 1—we find none that consistently differentiated between the high and low performing EdExAG groups in both years.

For Cohort 6, in Spring 1972, one positive outlier group and one negative group had mean scores below national norms on both pretest and posttest; for the other outliers, mean pretest and posttest scores were above the norms. The key variables for the Spring 1972 Reading Total analysis are presented in Table V-H-3 in Appendix V-H. Average yearly gains for students in the negative outlier groups ranged from .2 to .4 grade equivalents as compared with average gains of .7 to 1.2 grade equivalents for the positive outlier groups.

As a group, the 1972 positive outliers showed little difference from the negative outliers in terms of average student SES and proportion of girls. The positive and the negative groups also both had a high proportion of white students except for one negative outlier which had only 39 percent. Considering Reading Interest, the students in the positive outlier groups were generally lower in 1971, but in 1972 the group averages were about the same.

In the Spring 1973 analysis, the mean pretest and posttest scores were above national norms for all positive and all negative outliers (see Table V-H-4 in Appendix V-H). A gross indication of the differences in growth between the negative and positive outlier groups is indicated by the fact that average gains in the negative outlier groups were equal to or less than .3 of a grade equivalent between Spring 1972 and Spring 1973, whereas average gains in the positive outliers ranged between .8 and 1.0 grade equivalents.

Looking at the characteristics of the students in both sets of 1973 outliers, we found that all but one group averaged above the mean SES of our sample; the exception was a positive outlier group. We also found that the positive outliers had a generally higher proportion of white students, but that positive and negative groups were relatively equal as to the proportion of girls. There were no major differences between groups on Reading Interest.

As for treatment variables, once again there appears to be no clear evidence of treatment characteristics that distinguish high performing groups from low performing groups over the two years. Although in 1972 the positive outliers displayed a generally higher Level of Innovation, greater Degree of Individualization, and slightly more minutes per day on language arts instruction, in 1973 the positive outliers tended to be no different with respect to Number of Minutes per Day but somewhat less innovative and less individualized than the negative outliers. Also, in 1972 the high performing EdExAG groups generally scored higher on many of the individual scales of innovation; however, in 1973 these differences were absent, or in a few cases reversed. Group differences on the teacher variables were generally nonexistent for the Cohort 6 outliers across the two analysis years.

Considering the other profile variables listed in Appendix IV-H, once again none consistently differentiated between the positive and negative outliers across the two years.

To summarize, a search was made for possible patterns across all six outlier group analyses (i.e., the Spring 1972 and Spring 1973 analyses for each of the three cohorts). The results, at best, are mixed for the two superscales, Level of Innovation and Degree of Individualization. For two of the cohorts (Cohorts 1 and 6), the high performing groups in 1972 were characterized by both greater innovation and greater individualization, while in 1973 the high performing groups were exposed to less innovation and less individualization. For the other cohort (Cohort 4) the positive outliers in 1972 did spend more class time on language arts but differed little from the negative outliers on the dimensions of innovation and individualization. In 1973, the positive outliers in Cohort 4 were exposed to greater innovation, a higher degree of individualization, but about the same amount of instruction per day.

Considering teacher characteristics, there was some evidence of consistency in Cohort 1 with the positive outliers for both years generally having teachers with higher qualifications and lower SES background than the teachers of the negative outliers. However, these same findings did not occur consistently in the analyses for the other cohorts.

Examination of Unusual Student Growth Patterns in Reading Achievement

Our next approach was to investigate possible correlates of reading achievement based on those individual students who, during both the 1971-72 and 1972-73 school years, exhibited changes in reading achievement that were substantially greater or less than that predicted from their pretest and SES levels. In order to accomplish this, students in each cohort were identified whose residualized reading achievement posstest scores were more than one-half of a standard deviation away from the mean of the residuals in each of the two analysis years. Four different groups of unusual student reading achievement could be thus identified: (1) students who consistently achieved below expectation, (2) students who consistently achieved above expectation, (3) students who first overachieved, then underachieved relative to the rest of their cohort, and (4) students who first underachieved, then overachieved relative to their cohort. Because the growth patterns shown by the first two groups run counter to possible regression effects due to test unreliability and also because they represent distinctly opposite types of longitudinal achievement, our discussion will focus on these findings. Treatment differences for students who exhibited either of the two unusual but inconsistent patterns of achievement were also examined and these findings are briefly summarized at the end of this section.

The results pertaining to consistent underachievers and overachievers within Cohort 1 are shown in Table V-25. The two groups were close to one another in SES characteristics and ethnic composition, but there were proportionately more girls among the overachievers (59 percent) than among the underachievers (39 percent). Although it is not known how the Cohort 1 overachievers and underachievers compared prior to their participation in Project LONGSTEP, the overachievers were considerably higher in terms of their Reading Interest by the end of the third grade.

Both groups were above grade level in terms of CTMM Mental Age - Language at the time they were first tested, but the overachievers were about 9 months ahead of the underachievers. By the end of the third grade, however, the discrepancy between the two groups increased considerably. The Cohort 1 underachievers had an average CTBS Reading Total

TABLE V-25

Comparison of Key Analysis Variables for Students Exhibiting Unusual Longitudinal Growth Patterns In Two Consecutive School Years--Cohort 1, Reading Total Score Analysis

Variable		Consistent Underachievers		Consistent Overachievers	
Number of Students		57		56	
SES	$\bar{X}$	100.35		101.15	
	SD	10.40		9.48	
	g^1	0.19		-0.04	
Sex[2] (girl=1; boy=0)	$\bar{X}$	0.39		0.59	
Ethnic Group[2] (white=1; non-white=0)	$\bar{X}$	0.86		0.88	
		Analysis Year		Analysis Year	
		SP72	SP73	SP72	SP73
Pretest (Nat'l Norm MA=85 for SP72; Interpolated Nat'l Norm=321 for SP73)	$\bar{X}$	90.30	281.09	99.68	408.54
	SD	10.36	30.87	7.07	32.93
	g	0.11	0.33	-2.01	0.34
Posttest(Interpolated Nat'l Norm =321.5 for SP72; =375.5 for SP73)	$\bar{X}$	281.09	307.74	408.54	514.24
	SD	30.87	32.10	32.93	40.55
	g	0.33	-0.45	0.34	0.20
Posttest Residualized for Pretest and SES	$\bar{X}$	-53.10	-49.83	55.37	60.51
	SD	22.27	22.68	22.33	37.50
	g	-0.95	-0.90	0.76	0.96
Reading Interest at Time of Pretest	$\bar{X}$				
	SD				
	g				

(continued)

Note: Blank entries indicate no data collected.

[1] Index of skewness

$$g = \frac{\frac{\Sigma x^3}{N}}{\frac{\Sigma x^2}{N}\sqrt{\frac{\Sigma x^2}{N}}}$$

, where g = skewness index,

x = deviation of a score from $\bar{X}$, where $\bar{X}$ is that group's mean on variabl X, and

N = number of students.

[2] For binary coded variables the $\bar{X}$ is equal to the proportion of students coded with a "1."

TABLE V-25 (continued)

Variable		Consistent Underachievers SP72	Consistent Underachievers SP73	Consistent Overachievers SP72	Consistent Overachievers SP73
Reading Interest at Time of Posttest	$\bar{X}$		98.01		107.18
	SD		9.80		7.79
	g		-0.38		-0.38
Utilization of Objectives	$\bar{X}$	2.70	2.82	2.35	2.20
	SD	0.70	0.44	0.92	0.92
Individualization in Decision Making	$\bar{X}$	1.51	1.64	1.29	1.20
	SD	0.59	0.54	0.47	0.38
Teacher or Locally Developed Materials	$\bar{X}$	1.65	1.76	1.50	1.80
	SD	0.45	0.41	0.54	0.32
Individualization of Instructional Pace	$\bar{X}$	2.23	2.25	2.34	2.21
	SD	0.65	0.51	0.58	0.41
Scheduling Characteristics	$\bar{X}$	1.67	2.02	1.45	1.38
	SD	0.66	0.71	0.62	0.48
Use of Performance Agreements	$\bar{X}$	1.60	1.95	1.45	1.55
	SD	0.62	0.80	0.62	0.80
Classroom Group Organization	$\bar{X}$	2.04	2.37	2.21	2.20
	SD	0.77	0.48	0.70	0.40
Teaching Unit Composition	$\bar{X}$	2.11	2.25	1.74	1.85
	SD	0.71	0.65	0.71	0.76
Completeness of Instructional Package	$\bar{X}$	2.29	2.17	1.92	1.98
	SD	0.56	0.50	0.75	0.60
Utilization of Student Evaluation	$\bar{X}$	2.40	2.63	2.39	2.39
	SD	0.67	0.36	0.59	0.50
Level of Innovation	$\bar{X}$	20.18	21.86	18.63	18.76
	SD	4.24	2.69	4.81	3.61

(continued)

TABLE V-25 (continued)

Variable		Consistent Underachievers SP72	Consistent Underachievers SP73	Consistent Overachievers SP72	Consistent Overachievers SP73
Degree of Individualization	$\bar{X}$	7.74	8.47	7.41	7.35
Number of School Days per Year	$\bar{X}$	184.98	184.98	183.14	183.14
	SD	9.49	9.49	7.95	7.95
Number of Minutes per Day	$\bar{X}$	84.91	98.33	112.68	93.48
	SD	26.05	44.02	30.38	20.94
Teacher/Student Contact Ratio[3]	$\bar{X}$		4.35		3.93
	SD		1.51		1.30
Number of Teachers	$\bar{X}$	2.54	1.14	1.63	1.00
	SD	2.02	0.35	1.22	0.00
Teacher Age	$\bar{X}$	31.38	35.53	35.13	36.69
	SD	5.68	10.36	11.22	12.13
Teacher Socioeconomic Background	$\bar{X}$	102.18	103.67	100.07	100.08
	SD	7.35	11.61	6.50	9.51
Teaching Qualifications	$\bar{X}$	97.96	99.13	98.58	103.74
	SD	6.95	7.08	7.90	5.60

[3] Relative measure of teacher/student contact equal to the percentage of teacher time per individual student.

Expanded Scale Score of 281 at the end of the second grade (grade equivalent 1.9) and 308 at the end of the third grade (grade equivalent 2.4). The overachievers had an average reading achievement score of 409 at the end of second grade (grade equivalent 4.5) and an average reading achievement score of 514 at the end of the third grade (grade equivalent 7.3). By the end of the third grade, then, there was a net difference in reading achievement between these groups of about 4.9 grade equivalents.

When we compare the two groups on their overall Level of Innovation, the Cohort 1 overachievers are shown to have been exposed to less innovative language arts treatments (in terms of the scale Level of Innovation) for both years than the underachievers. When we compare the two groups on Degree of Individualization, the same pattern appears: the overachievers were exposed to less individualization than the underachievers. On individual scales of innovation, the Cohort 1 overachieving group shows this same tendency to have been exposed to a lesser amount of innovation than the underachievers. For example, the overachievers, on the average, were exposed to a somewhat lower level of Utilization of Objectives, Individualization in Decision Making, Scheduling Characteristics, Use of Performance Agreements, Teaching Unit Composition, and Completeness of Instructional Package.

Although the two groups were similar in the length of their school year, there were some notable differences between them in the Number of Minutes per Day spent on language arts. The Cohort 1 overachievers spent an average of nearly 30 minutes more per day during the second grade (SP72 analysis) than did the underachievers; in the third grade (SP73 analysis), the two groups were much closer in time spent on language arts. Although the initial amount of class time related to reading is not in itself a full explanation of treatment effectiveness, it is possible that this early, more intensive concentration on the teaching of reading and language arts in the second grade was highly beneficial to the overachieving group. The fact that the overachievers were exposed to almost 20 minutes per day less instruction in 1973 than in 1972 and yet, continued to show substantial achievement gains in 1973 suggests that early emphasis on language arts skills may produce lasting effects.

Considering the differences between these two groups in terms of teacher characteristics, we find that the Cohort 1 overachieving students were, on the average, taught more often by a smaller number of different language arts teachers. Their teachers, on the average, were typically older, had slightly lower SES backgrounds, and had higher scores on the Teaching Qualifications scale. Also, the overachievers had a slightly lower degree of teacher contact than the underachievers.

We turn now to a comparison of the Cohort 4 consistent underachievers and overachievers. As shown in Table V-26, they had about the same SES characteristics and ethnic composition. There tended to be more girls (54 percent) among the overachievers than among the underachievers (46 percent). The overachievers started out with a higher Reading Interest than the underachievers, and the magnitude of this difference in attitudes toward reading was generally maintained over the next two years.

The Cohort 4 consistent underachievers had an average CTBS Reading Total Expanded Scale Score of 440 (grade equivalent 5.2) at the end of the fourth grade, an average reading score of 436 (grade equivalent 5.1) at the end of the fifth grade, and an average reading score of 431 (grade equivalent 5.0) at the end of the sixth grade. Thus, the underachievers were above grade level as fourth-graders, but showed no growth (relative to CTBS norms) during the next two consecutive school years. The consistent overachievers, on the other hand, had an average reading score of 450 (grade equivalent 5.6) at the end of the fourth grade, an average reading score of 521 (grade equivalent 7.6) at the end of fifth grade, and an average reading score of 594 (grade equivalent 9.9) at the end of the sixth grade. The difference in mean reading achievement between these two groups was therefore about 4.9 grade equivalents by the end of the sixth grade.

A comparison between the Cohort 4 underachieving and overachieving groups on Level of Innovation shows that the overachievers were exposed, on the average, to a slightly lower Level of Innovation in the fifth grade (SP72) than were the underachievers. The same finding also applies to Degree of Individualization. In the sixth grade, however, the two groups appear to have been about equal on both these superscales.

TABLE V-26

Comparison of Key Analysis Variables for Students Exhibiting Unusual Longitudinal Growth Patterns In Two Consecutive School Years--Cohort 4, Reading Total Score Analysis

Variable		Consistent Underachievers		Consistent Overachievers	
Number of Students		69		96	
SES	$\bar{X}$	101.36		99.67	
	SD	9.73		9.40	
	g^1	0.12		0.19	
Sex^2 (girl=1; boy=0)	$\bar{X}$	0.46		0.54	
Ethnic $Group^2$ (white=1; non-white=0)	$\bar{X}$	0.90		0.92	
		Analysis Year		Analysis Year	
		SP72	SP73	SP72	SP73
Pretest (Nat'l Norm=419 for SP72; =457 for SP73)	$\bar{X}$	440.01	435.69	449.77	520.63
	SD	70.66	61.16	47.02	46.32
	g	0.13	0.19	-0.27	-0.02
Posttest(Nat'l Norm=457 for SP72; Interpolated Nat'l Norm=490.5 for SP73	$\bar{X}$	435.69	431.31	520.63	594.32
	SD	61.16	61.54	46.32	54.41
	g	0.19	0.12	-0.02	0.04
Posttest Residualized for Pretest and SES	$\bar{X}$	-36.95	-40.13	40.73	45.40
	SD	18.96	18.18	19.58	21.28
	g	-2.44	-2.03	1.63	1.25
Reading Interest at Time of Pretest	$\bar{X}$	98.33	96.85	101.69	100.24
	SD	10.66	11.66	9.99	9.61
	g	-0.23	0.16	-0.18	-0.32

Note: Blank entries indicate no data collected.

(continued)

[1] Index of skewness

$$g = \frac{\frac{\Sigma x^3}{N}}{\frac{\Sigma x^2}{N}\sqrt{\frac{\Sigma x^2}{N}}}$$

, where g = skewness index,

x = deviation of a score from $\bar{X}$, where $\bar{X}$ is that group's mean on variable X, and

N = number of students.

[2] For binary coded variables the $\bar{X}$ is equal to the proportion of students coded with a "1."

TABLE V-26 (continued)

Variable		Consistent Underachievers		Consistent Overachievers	
		SP72	SP73	SP72	SP73
Reading Interest at Time of Posttest	$\bar{X}$	96.85	100.29	100.24	103.14
	SD	11.66	9.74	9.61	9.11
	g	0.16	-0.12	-0.32	-0.09
Utilization of Objectives	$\bar{X}$	2.70	2.50	2.36	2.40
	SD	0.68	0.68	0.83	0.78
Individualization in Decision Making	$\bar{X}$	1.40	1.34	1.31	1.36
	SD	0.44	0.35	0.39	0.44
Teacher or Locally Developed Materials	$\bar{X}$	1.80	1.95	1.80	1.93
	SD	0.55	0.52	0.54	0.52
Individualization of Instructional Pace	$\bar{X}$	2.29	2.26	1.96	2.22
	SD	0.66	0.67	0.59	0.60
Scheduling Characteristics	$\bar{X}$	1.51	1.48	1.56	1.57
	SD	0.60	0.53	0.66	0.55
Use of Performance Agreements	$\bar{X}$	1.57	1.70	1.45	1.75
	SD	0.73	0.79	0.57	0.74
Classroom Group Organization	$\bar{X}$	1.99	2.12	1.89	2.11
	SD	0.81	0.63	0.72	0.54
Teaching Unit Composition	$\bar{X}$	1.85	1.88	1.76	1.75
	SD	0.64	0.69	0.67	0.65
Completeness of Instructional Package	$\bar{X}$	2.12	2.09	1.92	2.01
	SD	0.64	0.52	0.61	0.47
Utilization of Student Evaluation	$\bar{X}$	2.59	2.48	2.38	2.51
	SD	0.55	0.61	0.61	0.53
Level of Innovation	$\bar{X}$	19.81	19.80	18.38	19.62
	SD	4.06	3.28	4.00	3.41

(continued)

TABLE V-26 (continued)

Variable		Consistent Underachievers		Consistent Overachievers	
		SP72	SP73	SP72	SP73
Degree of Individualization	$\bar{X}$	7.85	7.78	7.10	7.84
Number of School Days per Year	$\bar{X}$	181.33	181.48	181.17	181.08
	SD	6.45	6.49	5.38	5.53
Number of Minutes per Day	$\bar{X}$	98.04	85.84	103.39	85.03
	SD	22.88	23.94	28.06	25.04
Teacher/Student Contact Ratio[3]	$\bar{X}$		3.37		3.37
	SD		0.60		0.57
Number of Teachers	$\bar{X}$	1.59	1.49	1.70	1.47
	SD	1.29	1.11	1.35	1.02
Teacher Age	$\bar{X}$	35.43	38.74	36.22	37.66
	SD	10.17	10.15	10.34	9.34
Teacher Socioeconomic Background	$\bar{X}$	99.27	98.53	100.13	100.15
	SD	10.26	8.64	11.17	9.90
Teaching Qualifications	$\bar{X}$	101.29	101.38	103.79	100.55
	SD	8.04	7.11	6.66	7.06

[3] Relative measure of teacher/student contact equal to the percentage of teacher time per individual student

The Cohort 4 overachievers in both years were also exposed to slightly lower Utilization of Objectives, Individualization of Instructional Pace, Teaching Unit Composition, and Completeness of Instructional Package than were the underachievers. These differences, however, were very small.

The two groups did not differ in terms of the length of their school year. In the fifth grade (SP72), the Cohort 4 overachievers did have a slightly greater Number of Minutes per Day of language arts instruction than the underachievers, but in the sixth grade (SP73), the average time spent on language arts was the same for both groups. The two groups were also very close to each other in terms of the teacher variables.

Contrary to what was found in the two previous cohorts, the Cohort 6 overachieving group had proportionately more boys (62 percent) than did the underachieving group (44 percent). The overachieving group also had a higher percentage of whites (94 percent) than the underachieving group (72 percent). The SES characteristics of the two groups however, were almost identical. In terms of their average Reading Interest the two groups were clearly different. In both years the consistent overachievers showed greater interest in reading than did the consistent underachievers (see Table V-27).

The Cohort 6 underachievers had an average CTBS Reading Total Expanded Scale Score of 523 (grade equivalent 7.6) at the end of the sixth grade. They had an average reading score of 504 (grade equivalent 7.1) at the end of the seventh grade, and an average reading score of 484 (grade equivalent 6.5) at the end of the eighth grade. In contrast, the Cohort 6 overachievers had an average reading score of 540 (grade equivalent 8.2) at the end of the sixth grade, an average reading score of 599 (grade equivalent 10.1) at the end of the seventh grade, and an average reading score of 661 (grade equivalent 12.3) at the end of the eighth grade. Thus, the average discrepancy between these groups of students had reached about 5.8 grade equivalents by the end of the eighth grade.

TABLE V-27

Comparison of Key Analysis Variables for Students Exhibiting Unusual Longitudinal Growth Patterns In Two Consecutive School Years--Cohort 6, Reading Total Score Analysis

Variable		Consistent Underachievers		Consistent Overachievers	
Number of Students		59		85	
SES	$\bar{X}$	101.19		100.68	
	SD	9.06		9.52	
	g^1	0.22		0.17	
Sex^2 (girl=1; boy=0)	$\bar{X}$	0.56		0.38	
Ethnic Group^2 (white=1; non-white=0)	$\bar{X}$	0.72		0.94	
		Analysis Year		Analysis Year	
		SP72	SP73	SP72	SP73
Pretest(Interpolated Nat'l Norm =490.5 for SP72; Nat'l Norm=526 for SP73)	$\bar{X}$	522.61	503.92	540.16	599.35
	SD	66.59	63.01	62.23	64.85
	g	0.47	0.63	-0.20	-0.07
Posttest(Nat'l Norm =526 for SP72; Interpolated Nat'l Norm =553.5 for SP73)	$\bar{X}$	503.92	484.33	599.35	660.89
	SD	63.01	65.71	64.85	67.01
	g	0.63	0.27	-0.07	-0.26
Posttest Residualized for Pretest and SES	$\bar{X}$	-36.60	-41.47	43.43	46.11
	SD	15.60	18.22	19.77	27.97
	g	-0.97	-1.22	1.36	2.30
Reading Interest at Time of Pretest	$\bar{X}$	99.58	97.56	104.13	102.31
	SD	10.32	9.51	9.60	9.13
	g	0.16	0.20	-0.38	0.00

Note: Blank entries indicate no data collected.

(continued)

[1] Index of skewness

$$g = \frac{\frac{\Sigma x^3}{N}}{\frac{\Sigma x^2}{N}\sqrt{\frac{\Sigma x^2}{N}}}$$

, where g = skewness index,

x = deviation of a score from $\bar{X}$, where $\bar{X}$ is that group's mean on variable X, and

N = number of students.

[2] For binary coded variables the $\bar{X}$ is equal to the proportion of students coded with a "1."

TABLE V-27 (continued)

Variable		Consistent Underachievers		Consistent Overachievers	
		SP72	SP73	SP72	SP73
Reading Interest at Time of Posttest	$\bar{X}$	97.56	95.73	102.31	102.71
	SD	9.51	9.84	9.13	9.64
	g	0.20	0.41	0.00	-0.44
Utilization of Objectives	$\bar{X}$	2.44	2.46	2.40	2.40
	SD	0.75	0.64	0.74	0.65
Individualization in Decision Making	$\bar{X}$	1.31	1.28	1.24	1.21
	SD	0.31	0.34	0.30	0.30
Teacher or Locally Developed Materials	$\bar{X}$	2.32	2.53	2.37	2.36
	SD	0.75	0.48	0.49	0.50
Individualization of Instructional Pace	$\bar{X}$	1.97	1.97	1.96	1.74
	SD	0.71	0.58	0.79	0.71
Scheduling Characteristics	$\bar{X}$	1.92	1.63	1.76	1.61
	SD	0.89	0.61	0.82	0.65
Use of Performance Agreements	$\bar{X}$	1.53	1.68	1.40	1.62
	SD	0.62	0.72	0.62	0.74
Classroom Group Organization	$\bar{X}$	1.90	1.81	2.09	1.98
	SD	0.48	0.43	0.29	0.22
Teaching Unit Composition	$\bar{X}$	1.61	1.64	1.59	1.61
	SD	0.55	0.60	0.59	0.66
Completeness of Instructional Package	$\bar{X}$	2.04	2.01	1.88	1.76
	SD	0.55	0.61	0.68	0.69
Utilization of Student Evaluation	$\bar{X}$	2.42	2.39	2.29	2.16
	SD	0.68	0.77	0.77	0.86
Level of Innovation	$\bar{X}$	19.45	19.39	18.99	18.45
	SD	4.97	4.35	4.70	4.67

(continued)

TABLE V-27 (continued)

Variable		Consistent Underachievers		Consistent Overachievers	
		SP72	SP73	SP72	SP73
Degree of Individualization	X̄	7.23	7.32	6.89	6.73
Number of School Days per Year	X̄	181.69	181.56	183.06	183.67
	SD	5.57	5.75	7.20	7.05
Number of Minutes per Day	X̄	57.12	51.69	54.00	49.65
	SD	9.22	7.40	10.20	7.27
Teacher/Student Contact Ratio[3]	X̄		3.53		3.36
	SD		0.52		0.51
Number of Teachers	X̄	2.69	3.12	2.16	2.61
	SD	2.26	2.44	1.84	2.36
Teacher Age	X̄	35.87	39.22	34.81	37.78
	SD	9.66	9.20	12.27	10.79
Teacher Socioeconomic Background	X̄	104.14	103.01	107.50	102.44
	SD	7.91	7.69	6.85	9.21
Teaching Qualifications	X̄	96.97	98.74	95.66	100.49
	SD	8.58	6.75	8.04	5.73

[3] Relative measure of teacher/student contact equal to the percentage of teacher time per individual student.

The Cohort 6 overachieving group was exposed to slightly lower Level of Innovation than the underachieving group for both years. The same trend was also true for Degree of Individualization. Differences between the two groups on specific dimensions of innovation were less clear. However, with the exception of Classroom Group Organization, the overachievers never scored higher than the underachievers on the same individual scale of innovation in both years.

The quantity of language arts instruction given to the two groups was roughly equal. The Cohort 6 overachievers had a slightly longer school year, but the underachievers spent slightly more minutes per day in language arts.

Teacher variables showed little difference between the two groups. One exception was the average number of teachers involved in language arts; on the average, the Cohort 6 overachievers had slightly fewer teachers than the underachievers in both the seventh and eighth grades.

Across all three cohorts, consistent overachievers and underachievers differed on Reading Interest. The overachievers consistently displayed a higher level of Reading Interest during both the pretest and posttest school years. However, we are reluctant to utilize Reading Interest as an explanatory variable because its relationship to the criterion is unclear--interest could have affected achievement, or achievement could have influenced attitude, or the relationship could have been reciprocal. To a somewhat lesser extent, the consistent overachievers tended to be involved in programs with a lower Level of Innovation and lower Degree of Individualization. The negative association of Level of Innovation and Degree of Individualization with unusual reading achievement was neither strong nor totally consistent, but it may still be considered a general finding. On the other hand, the absence of strong and consistent positive relationships clearly does not support the hypothesis that Level of Innovation or Degree of Individualization (as defined here) is positively related to dramatic gains in reading achievement, even for a select group of students who demonstrated such gains.

The analysis of consistently overachieving and underachieving students, however, did contain two additional findings that should be noted. First, the mean, standard deviation, and skewness of the SES distributions in the two groups in the analyses for each cohort were very similar. Although the possible statistical explanations for these results are complex (undoubtedly related to the lack of pretest/SES interactions and/or the lack of heteroscedasticity in the test/SES regressions), these findings do suggest that over- or underachievement, as defined here, was not differentially associated with student SES level. The second notable finding in these analyses was that Cohort 1 consistent overachievers on the average spent considerably more class time per day on language arts during the second grade than did underachievers--113 minutes compared to 84 minutes. In third grade, however, overachievers spent slightly less class time on language arts and yet, again demonstrated notable achievement gains. Although these results may be somewhat method dependent, they do suggest the interpretation that early emphasis on language arts skills may produce lasting effects.

As mentioned previously, an examination of inconsistent student growth patterns was also undertaken to see if the transition from overachievement to underachievement followed a pattern of association with predictor variables that was different from the pattern generated by the transition from underachievement to overachievement. In terms of overall Level of Innovation and Degree of Individualization, no clear patterns emerged. In all three cohorts, the overachievement-to-underachievement group, on the average, was exposed to about the same type of educational treatment from year to year. In Cohorts 1 and 4, the underachievement-to-overachievement students were exposed to programs with a slightly higher Level of Innovation and Degree of Individualization in 1973. In Cohort 6, however, the underachievement-to-overachievement students were members of slightly less innovative programs in 1973.

C. The Analysis of Language Achievement

An Overview of Language Achievement

The results of the language achievement analyses presented in this section were expected to be quite similar to those previously discussed for reading achievement because (1) the CTBS Language Total Score and Reading Total Score are known to be highly correlated (approximately .75 to .85 in the standardization samples; CTB/McGraw-Hill, 1970, pp. 61-62); and (2) the same "language arts" treatment and teacher data were associated with both outcome measures. The results obtained did, in fact, support this expectation. Differences between the language and reading analyses, to be sure, were present; however, none was judged to be unusually large and/or indicative of substantially different trends. This section, therefore, will only summarize those analyses which related treatment attributes to language achievement outcomes.

All of the analyses described in Chapter IV and discussed with respect to reading achievement were implemented on the six Language Total Score analysis samples. Tables similar to those found in the reading analysis sections and its associated appendices were generated and have been placed in Appendix V-I.

Correlations of key language arts treatment variables with that part of posttest that could not be linearly predicted from pretest and SES are shown in Appendix V-I, Table V-I-33. As with the reading analysis, many of these part correlations were statistically significant ($p < .05$), most with negative signs. There was a trend in Cohorts 1 and 6 for the SP72 correlations to be positive and the SP73 correlations, negative. In Cohort 4, this pattern tended to be reversed -- in Spring 1972 all statistically significant correlations were negative, in 1973 they were positive. Although the shift in the sign of the SP72 and SP73 correlations may have been due, in part, to test score unreliability and use of the same score as both a posttest and pretest measure, the large number of negative associations and complete lack of consistent signs across the two time points for a given cohort suggest that positive deviations from predicted posttest performance, if anything, tended to be associated with less innovative, more traditional treatment practices.

The commonality analyses of the Level of Innovation model (see Appendix V-I, Tables V-I-34 through V-I-39) and Degree of Individualization model (see Appendix V-I, Tables V-I-41 through V-I-46) showed that either of these superindices plus Number of Minutes per Day uniquely accounted for a notable proportion of variance in posttest language achievement in only one analysis sample, Cohort 1, SP73. Approximately 4 percent of the variation among posttest Language Total Scores was associated uniquely with the treatment attributes in this analysis sample. Since the overall multiple correlation squared of posttest with the treatment indicators ($\approx 2\%$) was less than this uniqueness, the same suppressor-like effect was undoubtedly occurring in these data as was discussed with reference to the reading results.

It should also be noted that the uniquenesses for the treatment variables in the language achievement growth models for Cohort 1, Spring 1973 were proportionately smaller than in the reading analyses. For example, the normalized uniquenesses for the treatment superindices in the language analysis were .07 and .07 (see Appendix V-I, Tables V-I-35 and V-I-42) for the Level of Innovation and Degree of Individualization models, while these same normalized coefficients in the reading analysis were .14 and .12, respectively (see Table V-15 in the text and Table V-G-2 in Appendix V-G). Thus, in relation to the reading achievement results, a somewhat smaller proportion of the predictable variation in posttest language performance can be uniquely attributed to these superscales.

The regression analysis results for the two growth models are found in Appendix V-I, Tables V-I-47 and V-I-48. Comparison of the patterns of statistically significant regression coefficients obtained in these analyses with those obtained for reading showed only a small number of discrepancies. The language achievement results with respect to Level of Innovation and Degree of Individualization demonstrated no trend for the regression coefficients for either of these superindices to be both statistically significant ($p < .05$) and of the same sign across all six analysis samples, or across each pair of analysis data sets for a given cohort. And in the one sample where a substantial uniqueness for Level of Innovation (or alternately, Degree of Individualization) plus Number of Minutes per Day was present (i.e., Cohort 1, SP73), the coefficients for both the superindex and the quantity of schooling measure were negative. The teacher index, Teaching Qualifications, had very few statistically significant regression coefficients and, in the two samples where it did receive a significant coefficient, one was negative (and a possible suppressor variable) and one was positive.

In summary, the small proportion of variance in the language posttest performance accounted for by treatment variables--and the lack of consistent positive findings when the criterion or the predictors were adjusted for pretest and SES differences--clearly do not support the hypothesis that Level of Innovation or Degree of Individualization (as defined here) is related to increased language achievement.

Examination of Unusually Effective EdExAG Groups

Presented in this section is a summary of the analysis of the treatment attributes of the EdExAG groups in each cohort with the greatest positive and the greatest negative mean residual scores for the Spring 1972 posttest analyses and the Spring 1973 posttest analyses. The key analysis variables for each group are shown in Tables V-I-49 through V-I-54 in Appendix V-I.

In order to identify any possible trends in the language outlier group analysis, the six separate analyses (Spring 1972 and Spring 1973 for each of the three cohorts) were compared. As was true for the similar analyses for reading, the results for either superscale, Level of Innovation or Degree of Individualization, were obviously conflicting. For no cohort were the results consistent over the two years. Neither were the results consistent in any single year across the three cohorts. In comparison with the poorer performing groups, the high performing EdExAG groups were just as often characterized by a lower Level of Innovation and a lower Degree of Individualization as they were by a higher score on these superscales. The results for the measure of Teaching Qualifications, however, were not quite so contradictory. In the analyses where the positive outliers did differ from the negative outliers on this scale, the teachers of the high performing EdExAG groups tended to possess higher Teaching Qualifications.

Examination of Unusual Student Growth Patterns in Language Achievement

Students who had residualized CTBS Language Total Expanded Scale Scores for both of the analysis years were categorized according to their deviation from the mean of their cohort. Students who were equal to or more than one-half of a standard deviation below the mean for both years were defined as consistent underachievers in language arts, while students who were equal to or over one-half of a standard deviation above the mean for both years were defined as consistent overachievers. Students who had unusual but inconsistent patterns of language achievement will be mentioned later. The results for the Cohort 1 consistent underachievers and overachievers in language are shown in Table V-I-55 in Appendix V-I.

Although the two groups were very similar with respect to SES characteristics, there was a distinct difference between them as to their sex composition: there were proportionately about twice as many girls among the Cohort 1 overachievers (69 percent) as there were among the underachievers (38 percent). The overachievers also had proportionately fewer whites (83 percent) than the underachievers (91 percent).

The consistent underachievers in Cohort 1 had an average CTMM Mental Age-Language of 93 at the end of the first grade, and average CTBS Language Total Expanded Scale Scores of 281 (grade equivalent 2.1) at the end of the second grade and 309 (grade equivalent 2.5) at the end of the third grade. The consistent overachievers had a CTMM Mental Age-Language of 96 at the end of the first grade, and average CTBS Language Scale Scores of 405 (grade equivalent 4.3) at the end of the second grade and 509 (grade equivalent 7.0) at the end of the third grade. One can thus characterize the difference between these two groups in language arts achievement as about 4.5 grade equivalents by the end of the third grade.

There was no difference between the Cohort 1 underachievers and overachievers in language in terms of the Level of Innovation and Degree of Individualization that they were exposed to in the second grade (Cohort 1, SP72). In the third grade, however, the overachievers were lower on both superscales than were the underachievers.

Although the two groups had the same average length of school year, the overachievers had longer average daily exposure to language arts during the second grade, while the underachievers had longer average daily exposure during the third grade. The overachieving group tended to have older and slightly more qualified language arts teachers in both analysis years.

The results for the Cohort 4 consistent underachievers and overachievers in language arts are shown in Table V-I-56 in Appendix V-I. Although the SES distributions of the two groups were similar, there were proportionately more whites among the overachievers (91 percent) than there were among the underachievers (86 percent). There were also more girls (58 percent) among the overachievers than among the underachievers (33 percent).

Regarding Attitude toward Language Arts, it appears that the two groups were most different from one another in initial attitudes, but over time, they seem to come closer together.

The Cohort 4 consistent underachievers in language arts had a CTBS Language Total Expanded Scale Score mean of 420 at the end of the fourth grade (grade equivalent 4.7), 416 at the end of the fifth grade (grade equivalent 4.6), and 415 at the end of the sixth grade (grade equivalent 4.6). The corresponding means for the consistent overachievers were 462 at the end of the fourth grade (grade equivalent 5.7), 541 at the end of the fifth grade (grade equivalent 8.0), and 609 at the end of the sixth grade (grade equivalent 10.4). Thus, the overachieving group was about 5.8 grade equivalents ahead of the underachieving group in language arts skills by the end of the sixth grade.

The two groups were clearly different in terms of the Level of Innovation and Degree of Individualization to which they were exposed. In both analysis years, it was the overachieving group that was exposed to the more conventional methods of language arts treatment.

Both groups had the same length of school year. The overachievers spent slightly less time in language arts per day during 1972 than the underachievers, but in 1973 they averaged nine minutes per day more of language arts instruction than the underachievers.

Although the differences were small, in both anaylsis years the overachievers tended to be taught by a larger number of older and more qualified teachers than the underachievers.

We turn now to a comparison of the unusually consistent underachievers and overachievers in language arts for Cohort 6. Table V-I-57 in Appendix V-I shows that although the two groups were very similar in SES composition and reasonably similar in ethnic composition, there was a notable difference between them in terms of their sex composition: 71 percent of the overachievers were girls, while only 45 percent of the underachievers were girls. The overachievers also tended to have a more positive Attitude toward Language Arts than did the underachievers.

The Cohort 6 consistent underachievers in language started out with an average CTBS Language Total Expanded Scale Score of 534 (grade equivalent 7.7) at the end of the sixth grade, but then went to an average language achievement score of 497 (grade equivalent 6.7) at the end of the seventh grade, and 465 (grade equivalent 5.8) at the end of the eighth grade. In contrast, the consistent overachievers started with an average CTBS Language Total Expanded Scale Score of 545 (grade equivalent 8.2) and then rose to an average language achievement score of 601 (grade equivalent 10.1) at the end of the seventh grade, and 660 (grade equivalent 12.3) at the end of the eighth grade. The difference between these two groups was therefore about 6.5 grade equivalents by the end of the eighth grade.

Despite these large differences in achievement patterns, there was little that differentiated the two groups in terms of their educational treatments.

The two groups were also similar to one another in terms of the time spent on language arts--neither one showed a consistent advantage in length of school year or in Number of Minutes per Day spent on language arts. None of the teacher variables differentiated consistently between the two groups either.

In summarizing these findings across the three cohorts, the most striking association with achievement group membership was that of student sex. Girls were much more likely to have been overachievers than boys.

Unlike the reading analysis, individual treatment variables did not differentiate between the two types of consistent achievers. However, the average Level of Innovation and Degree of Individualization among the overachievers did tend to be slightly lower than among the underachievers.

As was true for reading, the Cohort 1 overachievers on the average spent more time per day in language arts activities during the second grade than did the underachievers. Since the consistent overachievers continued to show dramatic gains in the third grade even though they spent somewhat less class time on language arts activities (100 minutes per day in grade 2 as compared with 90 minutes per day during grade 3), it is conceivable that emphasis on language arts during grade 2 may have had effects on these students lasting beyond the second grade.

An examination of unusual but inconsistent achievers was also undertaken to see if there were shifts in the types of language treatments being given to the students that corresponded with shifts in their achievement levels. However, no patterns clearly indicating an association between a change in treatment and a change in achievement in the language area could be found.

D. The Analysis of Arithmetic Achievement

Descriptive Overview of the Analysis Samples

Pretest/posttest gains and national norms. The difference between each pair of pretest and posttest means shown in Table V-28 was computed and divided by the pooled standard deviation (see equation V-1) to obtain a relative measure of the gains in mathematics achievement demonstrated by these LONGSTEP samples during each treatment year. These mean differences and "standardized" gains are shown in Table V-29. Examination of this table shows that relative gains seem to be inversely related to grade level since gains in arithmetic achievement were substantially greater in Cohort 1 than in Cohorts 4 and 6. These findings, like those in the reading and language achievement analyses, imply that meaningful correlates of achievement are more likely to be identified in the data from younger students, simply because proportionately greater changes in achievement were occurring during the early elementary school years.

As shown in Table V-28 and in Figure V-2, except for the Cohort 1, SP73 samples, the LONGSTEP analysis sample means are very close to national norms. Only in one data set (Cohort 6, Spring 1973) is the LONGSTEP sample mean below the norm. Comparison of pretest and posttest means with published national norms reveals that in only one analysis sample (Cohort 1, SP73) did the average student demonstrate gains greater than those in the CTBS norm sample. Cohort 6 students, on the other hand, lost ground relative to national norms during both the 1971-72 and 1972-73 school years. These same trends were present when test/norm differences were expressed in units of the posttest standard deviation (see Appendix V-J).

As noted in the reading achievement analyses, trends such as these could have been influenced, to some degree, by the effects of

- deleting from analysis all those students with missing data with respect to the growth model variables, and
- deleting students with perfect or near perfect scores on pretest or posttest because they represented students for whom the test ceiling was undoubtedly too low.

TABLE V-28

Descriptive Summary of the Arithmetic Total Score Analysis Samples

Variable	Cohort 1 SP72		Cohort 1 SP73		Cohort 4 SP72		Cohort 4 SP73		Cohort 6 SP72		Cohort 6 SP73	
Grade (pretest year/ posttest year)	1/2		2/3		4/5		5/6		6/7		7/8	
Total Number of Students	752		741		1943		1964		1443		1505	
Number of EdExAG Groups	18		17		42		43		10		17	
Number of Different Schools/Sites	13/5		13/5		35/13		34/13		7/7		7/7	
Mean ($\overline{X}$) and Standard Deviation (SD) by Variable	$\overline{X}$	SD	$\overline{X}$	SD	$\overline{X}$	SD	$\overline{X}$	SD	$\overline{X}$	SD	$\overline{X}$	SD
Posttest-Arithmetic Total Score Residualized for Time of Testing	320.02	38.90	377.79	42.41	446.66	56.43	480.84	61.58	510.30	72.52	537.81	73.94
National Norm - Posttest[1]	303	-	358	-	440	-	476.0	-	509.5	-	545	-
Pretest-Arithmetic Total Score Residualized for Time of Testing	93.83[2]	12.12[2]	319.15	37.43	404.69	37.68	449.44	56.06	487.46	57.44	517.17	66.72
National Norm - Pretest[1]	85[3]	-	303	-	400.0	-	440	-	476	-	509.5	-
Socioeconomic Status	99.80	9.93	100.49	9.88	99.26	9.94	99.55	9.97	99.10	9.79	99.74	9.67
Utilization of Objectives	2.65	0.73	2.68	0.64	2.50	0.78	2.55	0.71	2.89	0.41	2.81	0.44
Individualization in Decision Making[4]	1.32	0.43	1.50	0.51	1.33	0.39	1.19	0.37	1.23	0.30	1.14	0.25
Teacher or Locally Developed Materials	1.48	0.42	1.82	0.46	1.79	0.52	1.91	0.60	2.60	0.60	2.56	0.50

(continued)

---

[1] The 50th percentile score in Expanded Scale Score units for testing conducted between month 6 and 9 for Form R (SP71,SP73) and Form Q (SP72) - from norms table found in the CTBS Examiner's Manual. Interpolated scale score norms have a decimal point.

[2] Based on CTMM Mental Age - Non-Language score from Spring 1971.

[3] According to the CTMM Examiner's Manual a grade placement of 1.7 (the approximate time of testing) corresponds with a chronological age of 85 months.

[4] One of the scales summed to form the scale Degree of Individualization.

TABLE V-28 (continued)

Variable	Cohort 1 SP72		Cohort 1 SP73		Cohort 4 SP72		Cohort 4 SP73		Cohort 6 SP72		Cohort 6 SP73	
	$\overline{X}$	SD	$\overline{X}$	SD	$\overline{X}$	SD	$\overline{X}$	SD	$\overline{X}$	SD	$\overline{X}$	SD
Individualization of Instructional Pace[4]	2.26	0.68	2.24	0.43	2.05	0.64	2.27	0.76	2.30	0.46	2.38	0.71
Scheduling Characteristics	1.71	0.72	1.74	0.65	1.60	0.62	1.49	0.61	1.62	0.87	1.53	0.72
Use of Performance Agreements[4]	1.62	0.60	1.93	0.85	1.63	0.71	1.79	0.86	2.26	0.87	2.23	0.91
Classroom Group Organization	2.01	0.65	2.34	0.48	2.00	0.67	1.90	0.74	1.95	0.77	2.03	0.65
Teaching Unit Composition	1.88	0.78	2.22	0.66	1.75	0.72	1.83	0.64	2.28	0.73	1.60	0.67
Completeness of Instructional Package	2.21	0.65	2.15	0.59	2.04	0.58	2.16	0.60	2.67	0.42	2.35	0.66
Utilization of Student Evaluation[4]	2.40	0.60	2.49	0.50	2.48	0.55	2.50	0.54	2.82	0.47	2.54	0.72
Degree of Individualization[5]	7.59	1.67	8.17	1.45	7.49	1.63	7.76	1.74	8.61	0.87	8.30	1.87
Level of Innovation[6]	19.52	4.02	21.12	3.33	19.18	3.64	19.60	3.85	22.61	1.57	21.17	3.71
Number of Minutes per Day	52.89	8.45	50.86	8.14	52.72	8.22	51.56	10.06	52.24	8.83	51.92	9.54
Teacher Age	31.76	7.66	35.81	11.52	37.09	10.59	38.59	11.30	38.19	8.83	37.86	10.15
Teacher Socioeconomic Background	102.27	7.90	101.40	10.76	99.06	9.68	96.72	9.46	95.83	5.95	96.96	8.33
Teaching Qualifications	97.70	7.92	100.27	6.11	101.90	7.91	100.58	8.51	100.62	6.99	102.03	7.86

---

[4] One of the scales summed to form the scale Degree of Individualization.

[5] A superindex equal to the sum of the four scales noted in footnote 4.

[6] A superindex equal to the sum of the ten EdExAG scales, Utilization of Objectives through Utilization of Student Evaluation.

TABLE V-29

Mean Differences between Arithmetic Total Pretest and Posttest Scores and "Standardized" Gain

	Cohort 1[1] SP73	Cohort 4 SP72	Cohort 4 SP73	Cohort 6 SP72	Cohort 6 SP73
Mean Difference	58.64	41.97	39.40	22.84	20.64
"Standardized" Gain[2]	1.47	.87	.67	.35	.29

[1] The SP72 comparison is not included because the CTMM Mental Age-Non-Language was the pretest measure.

[2] From equation V-1.

The discussion in the reading achievement section showed that attrition (the main reason for missing data) effects and maximum test topout effects would more than likely cause gains to be overestimated, not underestimated by the students included in the analysis samples. It is likely that the same arguments are valid here also, since (1) the effects of missing data (see Chapter IV) were probably similar in both reading and these arithmetic analyses and (2) the maximum shifts in mean test scores due to the deletion of students with perfect or near perfect test scores were somewhat greater in the arithmetic analyses (compare Appendices V-K and V-B).

Therefore, it seems most reasonable to conclude that

- substantial gains in arithmetic achievement, based on comparisons among pretest, posttest and norms, were not present in these LONGSTEP samples;
- the greatest gains occurred in the Cohort 1, SP73 analysis sample; and
- there was a trend for the gains that were present to be inversely related to grade level.

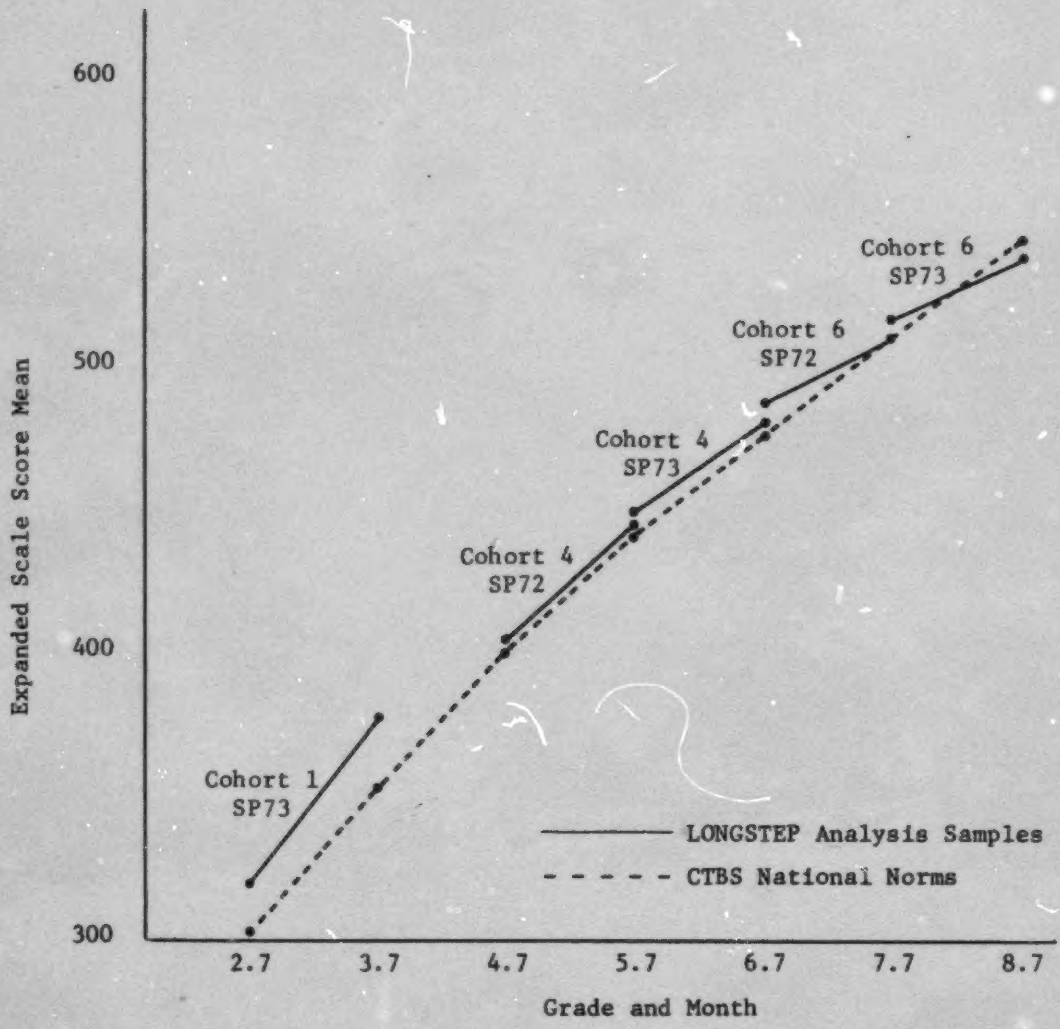


Figure [illegible]-2. Pretest and posttest means and published national norms (50th percentile for spring testing) - Arithmetic Total Score analyses.

Treatment attributes. The treatment variable means presented in Table V-28 show that, in general, students in all three analysis cohorts were members of educational treatment groups that ranked fairly high with respect to

- Utilization of Objectives
- Individualization of Instructional Pace
- Completeness of Instructional Package
- Utilization of Student Evaluation.

All analysis samples had a mean scale score that was above the theoretical midpoint (2.0) for each of these EdExAG variables. However, there is sufficient variation in the mean treatment attributes shown in Table V-28 to be able to conclude that students in different cohorts, on the average, were exposed to slightly different configurations of treatment components, different at least in terms of these key analysis variables. The one generalization that is somewhat apparent is that Cohort 6 students were involved in treatments that were slightly more innovative, in terms of overall Level of Innovation, Degree of Individualization, and the specific EdExAG/treatment scales themselves, especially during the 1971-72 school year. There were, however, no meaningful differences among the analysis samples with respect to the amount of time in each school day that was devoted to arithmetic instruction (i.e., the Number of Minutes per Day scale).

Students in the different analysis samples also were exposed to teachers who did not differ dramatically with respect to Age, SES Background or Teaching Qualifications. There were, however, two slight trends regarding average teacher differences. First, older students seemed to be exposed to older, more experienced teachers. For example, the average Age and Teaching Qualifications of the teachers in the second grade (Cohort 1, Spring 1972 posttest sample) were noticeably lower than in the other analysis samples. Second, younger students seem to have had math teachers whose SES Background was higher than that for teachers of older students.

Relationships among Treatments, Pretest and SES

Many of the analyses implemented on the arithmetic/mathematics data sets included statistical adjustments to "correct for" differences among EdExAG groups with respect to pretreatment variation in achievement level (i.e., pretest) and SES. Although such procedures can sometimes lead to misleading conclusions (e.g., "Lord's Paradox," Lord, 1967), they were necessary with the Project LONGSTEP data simply because EdExAG/treatment groups which differed on mean pretest or SES could be expected to differ on the posttest even when the various treatment approaches or teacher variables had no differential effects at all. It was valuable, therefore, to assess the extent to which different kinds of students were exposed to different treatments so that the impact of the statistical adjustments utilized at later stages might be more clearly understood. Commonality analysis with SES or pretest as the criterion variable and the zero-order correlations of the key analysis variables with SES and pretest were used to examine these relationships.

The commonality analyses of SES for the three-predictor-set model including pretest, the EdExAG dummy codes, and the three key teacher variables are contained in Appendix V-L; the proportions of SES variance attributable to EdExAG group membership [i.e., $R^2_{SES}(D_j)$] have also been placed in Table V-30. As shown in this table, on the average, about one-third to one-half of the variation in SES occurred among EdExAG groups [see $R^2_{SES}(D_j)$]. The among-group variance in SES, however, appeared to be proportionately larger in Cohort 4 than in either Cohort 1 or 6.

Table V-30 also shows the proportions of variance in pretest that were attributable to (1) EdExAG group differences [see $R^2_{pretest}(D_j)$]; (2) EdExAG group differences independent of concomitant differences with respect to SES and the teacher variables [i.e., the uniquenesses for pretest, $U_{pretest}(D_j)$]; and (3) SES variation within groups that was independent of teacher variable differences [$U_{pretest}(SES)$]. The complete commonality analysis results for pretest predicted from SES, the EdExAG group dummy codes, and the three key teacher variables are shown in Appendix V-M. One pretest trend which is apparent in Table V-30 is that EdExAG group differences with respect to pretest [$R^2_{pretest}(D_j)$] were proportionately

TABLE V-30

Proportions of SES and Pretest Variance Attributable to EdExAG Group Membership

	Cohort 1 SP72	Cohort 1 SP73	Cohort 4 SP72	Cohort 4 SP73	Cohort 6 SP72	Cohort 6 SP73
SES[1]						
$R^2_{SES}(D_j)$	.38	.41	.50	.53	.30	.32
Pretest						
$R^2_{pretest}(D_j)$	.08	.13	.13	.20	.10	.23
$U_{pretest}(D_j)$[2]	.08	.07	.09	.14	.05	.17
$U_{pretest}(SES)$[3]	.04	.05	.04	.03	.03	.02

[1] Among-group differences with respect to SES that were independent of concomitant differences in pretest were not computed because of the theoretical primacy of SES; that is, it was reasonable to view SES as an antecedent of test performance but not vice versa.

[2] This commonality analysis uniqueness is equal to the proportion of variance in pretest that is attributable to EdExAG group membership after group differences with respect to SES and the teacher variables are taken into account.

[3] This uniqueness is equal to the proportion of variance in pretest that is attributable to SES, independent of group differences on SES and all teacher variable differences; that is, it is a within-groups uniqueness.

greater for the SP73 analyses than for the SP72 analyses. The group membership uniquenesses [$U_{pretest}(D_j)$] indicate that these differences were not attributable to SES or to teacher variable differences among groups, at least for Cohorts 4 and 6. Since a higher proportion of among-group variance implies a lower proportion of within-group variance, on the average, the students within the SP73 EdExAG groups (for Cohorts 4 and 6) were more similar to each other, with respect to pretest level, than were the students in the SP72 EdExAG groups. The most likely explanation for this finding, however, may not be that there was a general tendency across the entire sample for each of the SP73 EdExAG groups to be more different with respect to mean initial arithmetic achievement level than the groups present in the SP72 analysis samples. Rather, these results were undoubtedly due, in part, to the fact that a substantially larger number of EdExAG groups were identified in Site 12 during the 1972-73 school year than during the 1971-72 school year (see Appendix IV-F, Tables 13 through 18), and that these groups not only received different treatments but were also fairly different with respect to mean initial achievement level in arithmetic.

The last row in Table V-30 [see $U_{pretest}(SES)$] shows that the within-group relationship between pretest and SES was small in all analysis samples.

Tables V-31 and V-32 contain the intercorrelations of each key analysis variable with SES and with pretest for each of the six analysis samples. Inspection of these correlations shows that there were a great many statistically significant relationships between student pretest and SES level and the kinds of treatments or teachers these students had during the 1971-72 and 1972-73 school years. There were no instances of correlations of SES with a given treatment or teacher attribute being both statistically significant and of the same sign across the three analysis cohorts. Table V-31 shows that, relative to the higher SES students in a given LONGSTEP analysis sample, those with lower SES tended to be members of programs that allocated a greater Number of Minutes per Day of class time to arithmetic/mathematics instruction. There also was a somewhat less notable trend across analysis samples for lower SES students to be exposed to programs characterized by a lower Level of Innovation and lower Degree of Individualization.

TABLE V-31

Correlations of SES with the Key Analysis Variables - Arithmetic Total Score Analysis

Level of Correlation	Variable	Cohort 1 SP72	Cohort 1 SP73	Cohort 4 SP72	Cohort 4 SP73	Cohort 6 SP72	Cohort 6 SP73
Student/ Student	Pretest	.169	.324	.294	.300	.295	.273
	Teacher Age	-.006	-.036	-.249	-.182	-.033	.020
	Teacher Socio-economic Background	-.028	.055	.340	.349	.059	.057
	Teaching Qualifications	.152	.173	-.171	-.037	.013	-.034
EdExAG/ Student	Utilization of Objectives	.303	.293	.104	.050	.009	.112
	Individualization in Decision Making	.039	.055	.036	.015	-.160	.084
	Teacher or Locally Developed Materials	.189	-.078	.067	-.111	-.127	-.192
	Individualization of Instructional Pace	.123	.229	-.078	-.154	-.286	-.127
	Scheduling Characteristics	-.038	-.059	.070	.020	.441	.488
	Use of Performance Agreements	-.044	-.086	-.087	-.005	-.051	.217
	Classroom Group Organization	-.110	-.130	-.117	-.059	.186	.055
	Teaching Unit Composition	.282	.176	.089	-.124	-.356	-.017
	Completeness of Instructional Package	.215	.285	-.031	-.201	-.402	-.142

(continued)

TABLE V-31 (continued)

Level of Correlation	Variable	Cohort 1 SP72	Cohort 1 SP73	Cohort 4 SP72	Cohort 4 SP73	Cohort 6 SP72	Cohort 6 SP73
EdExAG/ Student (cont.)	Utilization of Student Evaluation	.257	.280	.180	-.037	.058	.130
	Degree of Individualization	.136	.134	.001	-.079	-.227	.119
	Level of Innovation	.196	.159	.035	-.104	-.108	.123
	Number of Minutes per Day	-.101	.134	-.053	-.126	-.172	-.298
$r(p < .05)$ - Student/Student Level[1]		.071	.072	.044	.044	.051	.050
$r(p < .05)$ - EdExAG/Student Level[2]		.056	.055	.031	.030	.043	.042

Note: For convenience, all coefficients in this table were truncated to three decimal places--see Appendix V-N for correlations with five-place accurancy.

[1] Absolute value of the coefficient required for a correlation between two student-level variables to be statistically significant ($\alpha \leqq .05$, two-tailed). Due to the large number of degrees of freedom, the critical ratio $z = r/\sigma_r$ where $\sigma_r = 1/\sqrt{N}$ was used.

[2] Absolute value of the coefficient required for a correlation between one student-level variable (SES) and an EdExAG-level variable to be statistically significant ($\alpha \leqq .05$, two-tailed)--see equation IV-3 for the computing formula.

TABLE V-32

Correlations of Pretest with the Key Analysis Variables -
Arithmetic Total Score Analysis

Level of Correlation	Variable	Cohort 1 SP72	Cohort 1 SP73	Cohort 4 SP72	Cohort 4 SP73	Cohort 6 SP72	Cohort 6 SP73
Student/ Student	Socioeconomic Background	.169	.324	.294	.300	.295	.273
	Teacher Age	.040	.051	-.085	-.018	-.061	.067
	Teacher Socioeconomic Background	.027	.004	.128	.126	-.032	-.068
	Teaching Qualifications	-.033	.136	-.006	.083	-.031	.121
EdExAG/ Student	Utilization of Objectives	.069	.151	-.080	-.062	-.012	-.089
	Individualization in Decision Making	-.016	.013	-.026	-.053	-.064	-.142
	Teacher or Locally Developed Materials	-.051	-.025	.007	-.078	.028	-.158
	Individualization of Instructional Pace	.036	.078	-.135	-.249	-.221	-.331
	Scheduling Characteristics	.055	-.026	-.034	-.120	.256	.208
	Use of Performance Agreements	.025	.034	-.036	-.135	.118	-.066
	Classroom Group Organization	.144	.025	-.088	-.190	.076	-.023
	Teaching Unit Composition	.044	.082	-.069	-.205	-.175	-.266
	Completeness of Instructional Package	-.089	.077	-.151	-.248	-.218	-.311

(continued)

TABLE V-32 (continued)

Level of Correlation	Variable	Cohort 1 SP72	Cohort 1 SP73	Cohort 4 SP72	Cohort 4 SP73	Cohort 6 SP72	Cohort 6 SP73
EdExAG/ Student (cont.)	Utilization of Student Evaluation	.029	.134	.063	-.007	.030	-.035
	Degree of Individualization	.030	.094	-.053	-.190	-.005	-.192
	Level of Innovation	.047	.095	-.100	-.238	.043	-.196
	Number of Minutes per Day	.179	.120	-.061	-.002	-.142	-.114
r(p < .05) - Student/Student Level[1]		.071	.072	.044	.044	.051	.050
r(p < .05) - EdExAG/Student Level[2]		.069	.068	.042	.040	.049	.045

Note: For convenience, all coefficients in this table were truncated to three decimal places—see Appendix V-N for correlations with five-place accuracy.

[1] Absolute value of the coefficient required for a correlation between two student-level variables to be statistically significant ($\alpha \leqq .05$, two-tailed). Due to the large number of degrees of freedom, the critical ratio $z = r/\sigma_r$ where $\sigma_r = 1/\sqrt{N}$ was used.

[2] Absolute value of the coefficient required for a correlation between one student-level variable (pretest) and an EdExAG-level variable to be statistically significant ($\alpha \leqq .05$, two-tailed)—see equation IV-3 for the computing formula.

As with the SES correlations, the correlation patterns of pretest with the key treatment and teacher variables were not similar across all analysis cohorts. There was, however, a slight trend in Cohorts 4 and 6 for the SP72 correlations of the treatment and teacher variables with pretest to be statistically significant and negative. This tendency was not as apparent in the Cohort 1 samples where (1) very few statistically significant correlations were present in the SP72 analysis and (2) the statistically significant correlations were, with one exception, positive. The same trend--that is, for the correlations with pretest to be negative in Cohorts 4 and 6 and positive in Cohort 1--also tended to be present for Level of Innovation, Degree of Individualization and Number of Minutes per Day.

Table V-32 also shows that, for each separate cohort, there were more statistically significant correlations of these key analysis variables with pretest in the SP73 analyses than in the SP72 analyses. Although it is tempting to speculate that this may represent some kind of treatment impact (since the SP72 posttest becomes the pretest in the SP73 pretest/posttest analyses), it also seems likely that the larger number of significant correlations in the SP73 analyses (at least for Cohorts 4 and 6) are due to (1) a larger number of EdExAG groups in Site 12 during the 1972-73 school year compared to the 1971-72 school year, and (2) some form of ability grouping and implementation of different treatment approaches in these groups.

In summary, while there were many statistically significant correlations between pretest level and the attributes of students' educational treatment and teachers in arithmetic, the lack of similar signs for these correlations across all cohorts suggests that the pattern of pretest/treatment confounding differed by cohort. The SES correlations also varied across cohorts but not in the same way as the pretest correlations.

The intercorrelations among the EdExAG scales themselves (see Appendix V-N) were consistently positive and were moderate to high. It should be remembered, however, that the sampling error of these correlations is large because each coefficient is based not on the total number of students but rather on the number of EdExAG groups, each weighted by the number of students in the particular EdExAG groups. As discussed in Appendix IV-E, a somewhat more appropriate measure of the sample size to be used in testing the statistical significance of these weighted correlations is provided by the "effective sample size." Since the effective sample size is always less than the number of EdExAG groups when group _n_'s are unequal and since the number of EdExAG groups in these samples is relatively small (see Table V-28), only the largest of the correlations among the EdExAG scales reached statistical significance ($p < .05$). One trend shown in these correlations is the absence of significant associations between Number of Minutes per Day and any of the treatment scales, Level of Innovation, or Degree of Individualization.

Intercorrelations among the three key teacher scales (see Appendix V-N) were similar across all six analysis data sets. Teaching Qualifications was correlated positively with Teacher Age and negatively with Teacher Socioeconomic Background. Teacher Age was negatively correlated with Teacher Socioeconomic Background. Since years of experience is a major component of the Teaching Qualifications scale, it is not surprising that this scale was directly related to Teacher Age. The negative correlations with Teacher Socioeconomic Background, on the other hand, probably reflect the effects of the general increase in educational level in the United States in recent years.

The various analyses which examined the relationships between student achievement level in the previous year (i.e., pretest) and student SES have more or less shown that, on the average, schools with different kinds of students were implementing treatments whose basic underlying components were also systematically different. This does not necessarily imply that different educational approaches, as defined by the key variables considered here, were selectively applied to specific students within a school because of their SES or pretest levels. Although this explanation may have some validity with reference to the analysis of pretest, it seems more reasonable to conclude that sites or schools which were, on the average, composed of low SES students (i.e., low relative to all the students in a given LONGSTEP analysis sample) were implementing programs that emphasized practices that were somewhat different from those utilized in possibly more affluent schools. As noted in the analyses for reading achievement, the changes in sign of the SP72 and SP73 correlations between the treatment variables and pretest or SES probably reflect changes in treatments implemented between school years.

Relationships of Key Analysis Variables with Posttest Arithmetic Performance

Before examining the correlations between the key analysis variables and posttest, each posttest measure was subjected to a commonality and levels analysis so that the overlapping predictive relationships in a four-predictor model containing SES, pretest, EdExAG/treatment group membership and the three teacher scales (Age, SES Background and Teaching Qualifications) could be assessed. The next section contains the results of these analyses for the six analysis samples examined. Following that is a summary of the results for those readers who would rather not review these technical details.

Technical discussion of EdExAG group commonality analyses. As noted previously, for those readers who are unfamiliar with the general format of the commonality analysis tables used in this report, Appendix V-F provides an overview of how these tables may be read. Technical issues regarding the use of commonality analysis with group-level predictors (such as a set of dummy variables encoding mutually exclusive group membership) are discussed in Appendix IV-E.

Tables V-33 through V-38 show the commonality analyses which index (1) the proportions of arithmetic posttest variance that could be uniquely related to SES, pretest, a series of dummy variables encoding EdExAG group membership and the three key teacher variables, and (2) the second- and third-order commonality coefficients which provide a relative measure of the degree of predictive overlap of the variables in this particular prediction model. To facilitate comparisons of results across analysis samples, each commonality coefficient was divided by the total proportion of criterion variance that could be attributed to all four predictor sets. Dividing each coefficient by $R^2(S,P,D,T)$, the overall multiple correlation squared, produced what are called "normalized" commonality coefficients

TABLE V-33

Commonality Coefficients* for the Regression of Posttest on SES (S), Pretest (P), EdExAG Dummy Codes (D) and Teacher Variables (T) -- Cohort 1, Spring 1972, Arithmetic Total Score Analysis

	SES(S)	Pretest (P)	EdExAG Dummy Codes(D)	Teacher Variables(T)
$U(X_i)$	.03518 (.10739)	.11233 (.34289)	.07574 (.23121)	.01084 (.03309)
C(S,P)	.03436 (.10489)	.03436 (.10489)		
C(S,D)	.03074 (.09384)		.03074 (.09384)	
C(S,T)	.00083 (.00253)			.00083 (.00253)
C(P,D)		.01284 (.03918)	.01284 (.03918)	
C(P,T)		.00011 (.00033)		.00011 (.00033)
C(D,T)			.00427 (.01305)	.00427 (.01305)
C(S,P,D)	.00980 (.02993)	.00980 (.02993)	.00980 (.02993)	
C(S,P,T)	.00036 (.00110)	.00036 (.00110)		.00036 (.00110)
C(S,D,T)	.01134 (.03462)		.01134 (.03462)	.01134 (.03462)
C(P,D,T)		-.00617 (-.01883)	-.00617 (-.01883)	-.00617 (-.01883)
C(S,P,D,T)	-.00499 (-.01523)	-.00499 (-.01523)	-.00499 (-.01523)	-.00499 (-.01523)
$R^2(X_i)$	.11763	.15864	.13358	.01659

$R^2(S,P,D,T) = .32758$

---

*Below each commonality coefficient and in parentheses is the normalized coefficient; that is, the commonality coefficient divided by $R^2(S,P,D,T)$.

TABLE V-34

Commonality Coefficients* for the Regression of Posttest on SES (S), Pretest (P), EdExAG Dummy Codes (D) and Teacher Variables (T) -- Cohort 1, Spring 1973, Arithmetic Total Score Analysis

	SES(S)	Pretest (P)	EdExAG Dummy Codes(D)	Teacher Variables(T)
$U(X_i)$	.00785 (.01235)	.43975 (.69203)	.10350 (.16288)	.01719 (.02706)
C(S,P)	.05941 (.09350)	.05941 (.09350)		
C(S,D)	-.00771 (-.01213)		-.00771 (-.01213)	
C(S,T)	-.00100 (-.00158)			-.00100 (-.00158)
C(P,D)		.01506 (.02371)	.01506 (.02371)	
C(P,T)		.01830 (.02880)		.01830 (.02880)
C(D,T)			-.00887 (-.01396)	-.00887 (-.01396)
C(S,P,D)	-.01723 (-.02712)	-.01723 (-.02712)	-.01723 (-.02712)	
C(S,P,T)	-.00069 (-.00109)	-.00069 (-.00109)		-.00069 (-.00109)
C(S,D,T)	.00087 (.00137)		.00087 (.00137)	.00087 (.00137)
C(P,D,T)		-.00416 (-.00654)	-.00416 (-.00654)	-.00416 (-.00654)
C(S,P,D,T)	.01318 (.02074)	.01318 (.02074)	.01318 (.02074)	.01318 (.02074)
$R^2(X_i)$	.05467	.52362	.09464	.03482

$R^2(S,P,D,T) = .63544$

---

*Below each commonality coefficient and in parentheses is the normalized coefficient; that is, the commonality coefficient divided by $R^2(S,P,D,T)$.

TABLE V-35

Commonality Coefficients* for the Regression of Posttest on SES (S), Pretest (P), EdExAG Dummy Codes (D) and Teacher Variables (T) -- Cohort 4, Spring 1972, Arithmetic Total Score Analysis

	SES(S)	Pretest (P)	EdExAG Dummy Codes(D)	Teacher Variables(T)
$U(X_1)$	.00182 (.00264)	.49147 (.71379)	.04986 (.07242)	.00064 (.00093)
C(S,P)	.04045 (.05875)	.04045 (.05875)		
C(S,D)	.00163 (.00237)		.00163 (.00237)	
C(S,T)	-.00000 (-.00000)			-.00000 (-.00000)
C(P,D)		.05193 (.07542)	.05193 (.07542)	
C(P,T)		.00487 (.00707)		.00487 (.00707)
C(D,T)			.00378 (.00549)	.00378 (.00549)
C(S,P,D)	.02564 (.03724)	.02564 (.03724)	.02564 (.03724)	
C(S,P,T)	.00008 (.00012)	.00008 (.00012)		.00008 (.00012)
C(S,D,T)	.00146 (.00211)		.00146 (.00211)	.00146 (.00211)
C(P,D,T)		-.00422 (-.00612)	-.00422 (-.00612)	-.00422 (-.00612)
C(S,P,D,T)	.01913 (.02778)	.01913 (.02778)	.01913 (.02778)	.01913 (.02778)
$R^2(X_i)$	.09021	.62935	.14920	.02574

$R^2(S,P,D,T) = .68854$

---

*Below each commonality coefficient and in parentheses is the normalized coefficient; that is, the commonality coefficient divided by $R^2(S,P,D,T)$.

TABLE V-36

Commonality Coefficients* for the Regression of Posttest on SES (S), Pretest (P), EdExAG Dummy Codes (D) and Teacher Variables (T) -- Cohort 4, Spring 1973, Arithmetic Total Score Analysis

	SES(S)	Pretest (P)	EdExAG Dummy Codes(D)	Teacher Variables(T)
$U(X_i)$	.00141 (.00199)	.49598 (.69734)	.03957 (.05563)	.00230 (.00323)
C(S,P)	.03116 (.04381)	.03116 (.04381)		
C(S,D)	.00187 (.00263)		.00187 (.00263)	
C(S,T)	.00006 (.00008)			.00006 (.00008)
C(P,D)		.07153 (.10057)	.07153 (.10057)	
C(P,T)		.00283 (.00398)		.00283 (.00398)
C(D,T)			-.00184 (-.00259)	-.00184 (-.00259)
C(S,P,D)	.03839 (.05398)	.03839 (.05398)	.03839 (.05398)	
C(S,P,T)	.00101 (.00142)	.00101 (.00142)		.00101 (.00142)
C(S,D,T)	.00139 (.00196)		.00139 (.00196)	.00139 (.00196)
C(P,D,T)		.00428 (.00602)	.00428 (.00602)	.00428 (.00602)
C(S,P,D,T)	.02131 (.02996)	.02131 (.02996)	.02131 (.02996)	.02131 (.02996)
$R^2(X_i)$	.09660	.66649	.17650	.03133

$R^2(S,P,D,T) = .71124$

---

*Below each commonality coefficient and in parentheses is the normalized coefficient; that is, the commonality coefficient divided by $R^2(S,P,D,T)$.

TABLE V-37

Commonality Coefficients* for the Regression of Posttest on SES (S), Pretest (P), EdExAG Dummy Codes (D) and Teacher Variables (T) -- Cohort 6, Spring 1972, Arithmetic Total Score Analysis

	SES(S)	Pretest (P)	EdExAG Dummy Codes(D)	Teacher Variables(T)
U(X_i)	.00200 (.00290)	.53812 (.78017)	.01860 (.02696)	.00201 (.00291)
C(S,P)	.03582 (.05193)	.03582 (.05193)		
C(S,D)	.00117 (.00170)		.00117 (.00170)	
C(S,T)	.00013 (.00019)			.00013 (.00019)
C(P,D)		.03591 (.05206)	.03591 (.05206)	
C(P,T)		.00648 (.00940)		.00648 (.00940)
C(D,T)			.00378 (.00548)	.00378 (.00548)
C(S,P,D)	.04896 (.07098)	.04896 (.07098)	.04896 (.07098)	
C(S,P,T)	.00032 (.00046)	.00032 (.00046)		.00032 (.00046)
C(S,D,T)	-.00070 (-.00101)		-.00070 (-.00101)	-.00070 (-.00101)
C(P,D,T)		.00093 (.00135)	.00093 (.00135)	.00093 (.00135)
C(S,P,D,T)	-.00379 (-.00549)	-.00379 (-.00549)	-.00379 (-.00549)	-.00379 (-.00549)
$R^2(X_i)$	.08392	.66276	.10486	.00917

R^2(S,P,D,T) = .68975

---

*Below each commonality coefficient and in parentheses is the normalized coefficient; that is, the commonality coefficient divided by R^2(S,P,D,T).

TABLE V-38

Commonality Coefficients* for the Regression of Posttest on SES (S), Pretest (P), EdExAG Dummy Codes (D) and Teacher Variables (T) -- Cohort 6, Spring 1973, Arithmetic Total Score Analysis

	SES(S)	Pretest (P)	EdExAG Dummy Codes(D)	Teacher Variables(T)
U(X_1)	.00177 (.00236)	.50716 (.67578)	.01477 (.01968)	.00036 (.00048)
C(S,P)	.02763 (.03682)	.02763 (.03682)		
C(S,D)	.00215 (.00286)		.00215 (.00286)	
C(S,T)	.00024 (.00031)			.00024 (.00031)
C(P,D)		.12171 (.16217)	.12171 (.16217)	
C(P,T)		.00698 (.00929)		.00698 (.00929)
C(D,T)			.00110 (.00147)	.00110 (.00147)
C(S,P,D)	.05792 (.07717)	.05792 (.07717)	.05792 (.07717)	
C(S,P,T)	.00452 (.00602)	.00452 (.00602)		.00452 (.00602)
C(S,D,T)	-.00068 (-.00090)		-.00068 (-.00090)	-.00068 (-.00090)
C(P,D,T)		.01390 (.01852)	.01390 (.01852)	.01390 (.01852)
C(S,P,D,T)	-.00904 (-.01205)	-.00904 (-.01205)	-.00904 (-.01205)	-.00904 (-.01205)
$R^2(X_1)$	.08450	.73077	.20182	.01737

$R^2(S,P,D,T) = .75048$

---

*Below each commonality coefficient and in parentheses is the normalized coefficient; that is, the commonality coefficient divided by $R^2(S,P,D,T)$.

These have the advantage of summing to 1.0, no matter what the total predictability of the regression model happens to be.

Tables V-33 through V-38 show that except for the Cohort 1, SP72 analysis sample about two-thirds to three-fourths of the variation in the CTBS Arithmetic Total Expanded Scale Score could be accounted for by this particular four-predictor-set model [see the $R^2(S,P,D,T)$]. These predictors, however, only accounted for about one-third of variance in the Cohort 1, SP72 criterion variable, a result undoubtedly due to the substantially lower pretest/posttest correlation in these data. This pretest measure was the CTMM Mental Age - Non-Language score. The lower reliability of the scores derived from the CTMM, as well as the fact that the CTMM was designed to be more of an ability test than an achievement test, probably accounts for this depressed pretest/posttest correlation. (See the earlier discussion of the use of the CTMM Mental Age-Language score as a pretest measure in the Cohort 1, SP72 analysis for reading achievement.)

Probably because the pretest/SES correlation in the Cohort 1, SP72 analysis sample ($r_{\text{pretest,SES}} = .17$; see Appendix V-N) was also lower than the same correlation in the other samples (e.g., $r_{\text{pretest,SES}} = .32$ for Cohort 1, SP73; see Appendix V-N), there was (1) slightly less predictive overlap of SES and pretest and (2) a tendency for the Cohort 1, SP72 uniqueness for SES to be proportionately greater than in the remaining five samples where the uniquenesses were almost equal to zero. As expected, the uniquenesses for pretest in this Cohort 1 data set were also noticeably less than in the other data sets. Last, because of the small correlation of SES with the CTMM pretest measure, the confounding between SES and group membership [C(S,D)] was larger for Cohort 1, SP72 than for the other samples, whereas the SES, pretest and group membership confounding [C(S,D)] tended to be slightly smaller.

The $R^2(X_1)$ value for the EdExAG dummy code predictors in each commonality table indexes the proportion of variation in posttest arithmetic achievement that could be accounted for by merely knowing the EdExAG group membership of each student in the analysis sample. These proportions ranged from a low of 9 percent (for Cohort 1, SP73) to a high of 20 percent (for Cohort 6, SP73). Although it is possible to think of these proportions as providing a measure of the magnitude of the differences (on the average, across all students in an analysis sample) among the EdExAG group means on posttest, it is also possible to interpret them as measures of the accuracy with which the average student's posttest score could be predicted in these data. Thus, a multiple correlation squared of .20 means that by merely knowing the group membership of each student, we can predict his/her posttest score by using the mean of the group of which he/she is a member as a predictor variable (instead of the overall mean computed across all students), and in so doing, increase the accuracy of prediction by 20 percent. (A more technical presentation of how this form of variance partitioning, sometimes called levels analysis, is accomplished via the use of dummy variables in regression analysis is presented in Appendix IV-E.)

The analyses of pretest and SES presented previously, however, showed that EdExAG groups differed from each other in terms of average SES or pretest level. And, groups which differ with respect to initial achievement level would also differ on posttest, even when the various treatments had no differential impact at all. Therefore, a more reasonable estimate of the magnitude of the posttest differences among EdExAG groups attributable to the influences of treatment is provided by the uniqueness $[U(X_1)]$ for the EdExAG dummy codes shown in Tables V-33 through V-38. These tables show that when EdExAG group differences with respect to SES, pretest and the three key teacher variables have been taken into account, the proportions of variation in arithmetic posttest achievement that can be attributed to EdExAG/treatment group membership decrease greatly. Nevertheless, the proportion of criterion variance that can be accounted for by unique EdExAG group differences is considerable for Cohort 1 (.08 and .10 for SP72 and SP73, respectively) and moderate (.05 and .04) for Cohort 4. Since the uniquenesses for Cohort 6 were smaller yet, these results suggest that treatment impact was inversely related to grade level in these LONGSTEP samples.

In one analysis data set (Cohort 1, SP73, in Table V-34), however, the uniqueness for EdExAG group membership was even larger than the square of the multiple correlation of the posttest with the EdExAG dummy codes alone. It is possible for a part correlation squared, such as a commonality analysis uniqueness, to be larger than a simple correlation squared. When this occurs, it is possible that some kind of "suppressor effect" is present in which the covariates (SES, pretest and the teacher variables in this case) are accounting for the variation in a predictor (such as the dummy variables encoding group membership) that is not related to the criterion variable and which attenuates the predictor/criterion relationship. The appearance of a fairly large number of negative second- and third-order commonality coefficients in Table V-34 does indicate that negative predictive overlap is present in this system of predictors--implying that when these confounded variables are included, rather than partitioned out, the relationship of group membership with posttest arithmetic performance is attenuated.

Only in the two Cohort 1 analysis samples were the uniquenesses for the three teacher variables greater than .01. Although accounting for between 1 and 2 percent of the variance in posttest does not suggest a very large relationship, these two uniquenesses are notable given the minimal extent to which the treatment measures are related to posttest in the LONGSTEP data. The small multiple correlations between the teacher variable predictors and posttest were primarily confounded with the SES, pretest and EdExAG group [C(S,P,D,T)] in these samples.

As noted previously, the uniquenesses for SES were very small in all samples except Cohort 1 in SP72. Although there was considerable confounding of EdExAG group membership and SES (see Appendix V-L), most of the predictive overlap of SES and group membership in Tables V-33 through V-38 occurred in conjunction with pretest [e.g., C(S,P,D)]. Since the uniqueness for SES is equal to the proportion of total posttest variance attributable to within-group variation in SES that is not also confounded with within-group variation in pretest and the teacher variables, these results demonstrate that SES differences among students in the treatment groups were not related to posttest performance to any meaningful degree.

Pretest, as expected, was highly related to posttest, except in the Cohort 1, SP72 sample where the CTMM Mental Age - Non-Language score was used as the pretest measure. With this exception, variation on pretest within the EdExAG groups in each sample accounted for about one-half of the posttest variance. Pretest was also the one measure in the model with which all the other measures were confounded (see the second-order commonalities involving pretest in Tables V-33 through V-38). Although pretest was confounded with EdExAG group membership in all samples, there was a trend in Cohorts 4 and 6 for the predictive overlap of these two variable sets [i.e., C(P,D)] to be greater in the SP73 analysis samples than in their respective SP72 samples. This was probably due to the fact that group differences with respect to mean pretest level (independent of concomitant SES differences) were relatively larger in the SP73 analysis samples for Cohorts 4 and 6 than in the SP72 samples (see Table V-30).

Summary of EdExAG group commonality analyses. Commonality analysis showed that after SES, pretest and teacher variable differences among groups were taken into account, EdExAG group differences accounted for (1) as much as 10 percent and as little as 1 percent of the posttest variance in these six analysis samples, and (2) less variance in the posttest for the older cohorts. These findings suggest that growth in arithmetic achievement is negatively related to age/grade level.

Two other findings, though not dramatic, also seem worthy of note. First, except for the Cohort 1, SP72 analysis sample in which the CTMM provided the pretest measure, the unique association of SES with posttest was almost nil—that is, once group and individual student differences with respect to pretest were taken into account, SES had very little relationship with posttest achievement. Second, the teacher variables, independent of student pretest, SES and EdExAG group differences, accounted for about 1 percent of the posttest variance in the two Cohort 1 analysis samples. Although not large, these uniquenesses were notable compared to the small uniquenesses for the teacher variables in the other samples.

Correlations with posttest. The correlations of posttest with each of the key analysis variables were examined next to see if posttest arithmetic achievement was systematically related to any of the particular treatment or teacher attributes of prime interest in these analyses.

Table V-39 contains these posttest correlations for the six analysis samples examined. In relation to Cohorts 4 and 6, the number of statistically significant correlations with posttest in the two Cohort 1 samples was low.

Lastly, it should be noted that our Level of Innovation superscale was significantly correlated with posttest in five data sets. Of these, four were negative. Similarly, our Degree of Individualization superscale exhibited significant correlations with posttest in only three analysis data sets. All three were negative. Number of Minutes per Day, the study's quantity of instruction index, was positively correlated with posttest in the Cohort 1 samples, but negatively in Cohorts 4 and 6.

These zero-order correlations with posttest, however, do not take account of the fact that different kinds of students were exposed to different levels of the treatment variables. It is possible, then, that posttest was correlated with a given treatment attribute not because of any "effects" a treatment had but because students exposed to different levels of treatment also began the school year at different achievement levels and actually showed no differential improvement by the end of the school year (i.e., by posttest time).

An individual student's performance on the posttest is highly related to his or her pretest score and SES. Since students enter a given school in a given school year with these characteristics, it seems worthwhile to determine what the relationship is between the treatments and that part of arithmetic achievement that cannot be predicted from initial achievement level and SES. As noted previously, adjusting posttest on the basis of pretest and SES involves a different educational and statistical model from that involved in adjusting predictors (e.g., EdExAG group membership or the EdExAG variables) for pretest and SES differences. The complex manner in which student background and treatment attributes are confounded in public schools such as those participating in Project LONGSTEP also contributed to the decision to employ this additional analytic approach and educational model in examining treatment/posttest associations.

TABLE V-39

Correlations of All Key Analysis Variables with Posttest - Arithmetic Total Score Analysis

Level of Correlation	Variable	Cohort 1 SP72	Cohort 1 SP73	Cohort 4 SP72	Cohort 4 SP73	Cohort 6 SP72	Cohort 6 SP73
Student/ Student	Socioeconomic Background	.342	.233	.300	.310	.289	.290
	Pretest	.398	.723	.793	.816	.814	.854
	Teacher Age	-.094	.089	-.065	-.039	.005	.056
	Teacher Socioeconomic Background	.045	-.026	.158	.144	-.057	-.087
	Teaching Qualifications	.009	.183	-.039	.054	.000	.121
EdExAG/ Student	Utilization of Objectives	.036	-.101	-.049	-.110	.089	-.128
	Individualization in Decision Making	-.067	-.090	-.046	-.045	-.133	-.132
	Teacher or Locally Developed Materials	.071	-.058	.033	-.110	.082	-.190
	Individualization of Instructional Pace	.042	.005	-.164	-.242	-.199	-.323
	Scheduling Characteristics	.034	-.157	-.000	-.089	.283	.200
	Use of Performance Agreements	-.023	-.047	-.036	-.102	.125	-.110
	Classroom Group Organization	-.083	-.042	-.065	-.127	.020	-.072
	Teaching Unit Composition	.007	-.086	.023	-.191	-.214	-.200

(continued)

TABLE V-39 (continued)

Level of Correlation	Variable	Cohort 1 SP72	Cohort 1 SP73	Cohort 4 SP72	Cohort 4 SP73	Cohort 6 SP72	Cohort 6 SP73
	Completeness of Instructional Package	.077	-.057	-.142	-.269	-.139	-.323
	Utilization of Student Evaluation	.008	-.027	.078	-.043	.115	-.075
	Degree of Individualization	-.005	-.067	-.064	-.180	.034	-.225
	Level of Innovation	.018	-.121	-.064	-.232	.103	-.222
	Number of Minutes per Day	.099	.120	-.057	-.013	-.148	-.160
$r(p < .05)$ - Student/Student Level[1]		.071	.072	.044	.044	.051	.050
$r(p < .05)$ - EdExAG/Student Level[2]		.067	.069	.041	.040	.048	.045

Note: For convenience, all coefficients in this table were truncated to three decimal places--see Appendix V-N for correlations with five-place accuracy.

[1] Absolute value of the coefficient required for a correlation between two student-level variables to be statistically significant ($\alpha \leqq .05$, two-tailed). Due to the large number of degrees of freedom, the critical ratio $z = r/\sigma_r$ where $\sigma_r = 1/\sqrt{N}$ was used.

[2] Absolute value of the coefficient required for a correlation between one student-level variable (Pretest) and an EdExAG-level variable to be statistically significant ($\alpha \leqq .05$, two-tailed)--see equation IV-3 for the computing formula.

Correlations of the key analysis treatment variables with that part of posttest achievement which could not be linearly predicted from pretest and SES level are presented in Table V-40. Examination of Table V-40 shows that a majority of the statistically significant correlations are negative, implying that (1) students who obtained posttest scores higher than those expected on the basis of their pretest and SES level tended to have been exposed to treatments which were relatively more traditional; while (2) students who achieved less than expected tended to be members of EdExAG groups that were relatively more innovative. Cohorts 1 and 6 generally had more statistically significant correlations than Cohort 4. If both the SP72 and SP73 correlations for each cohort are considered together, there was a trend for the pairs of statistically significant correlations to be (1) of the same sign and negative in both SP72 and SP73 for Cohort 1, and (2) positive in SP72 and negative in SP73 for Cohort 6.

The shift in signs across SP72 and SP73 Cohort 6 analysis sample correlations was so prevalent that project staff speculated how this trend might have been affected by loss of cases due to missing data and test ceiling effects. The possible impact of these factors was not to be estimated directly since (1) attrition and complete loss of posttest year data was the main source of missing data and (2) no additional test scores were available which could be used for assessing the extent of the test ceiling effects in the tests used by Project LONGSTEP. Given the nature of the intercorrelations among presence/absence of complete data, pretest, posttest and Level of Innovation in the Cohort 6 samples, it seems likely that if these students had been in the Cohort 6 analysis samples, gains would have been (1) less positive in SP72 and (2) less negative in SP73. In either case, gains probably would have more closely approximated national norms. More importantly, the impact of these two factors would not, in all likelihood, have affected the conclusions about the presence or absence of dramatic treatment effects in these samples.

In summary, it should be noted that our superscale, Level of Innovation, exhibited significant correlations with that part of posttest achievement that could not be predicted from pretest and SES in four analysis data sets. Three of the four were negative. Similarly, of the three

TABLE V-40

Correlations of the Key Treatment Variables with Posttest Residualized on the Basis of Pretest and SES - Arithmetic Total Score Analyses

Variable	Cohort 1 SP72	Cohort 1 SP73	Cohort 4 SP72	Cohort 4 SP73	Cohort 6 SP72	Cohort 6 SP73
Utilization of Objectives	-.084	-.305	.008	-.113	.171	-.118
Individualization in Decision Making	-.084	-.145	-.047	-.005	-.127	-.035
Teacher or Locally Developed Materials	.041	-.058	.037	-.070	.115	-.089
Individualization of Instructional Pace	-.006	-.073	-.089	-.058	-.012	-.073
Scheduling Characteristics	.029	-.200	.035	.007	.096	-.007
Use of Performance Agreements	-.022	-.105	-.003	.010	.057	-.132
Classroom Group Organization	-.117	-.088	.018	.049	-.087	-.109
Teaching Unit Composition	-.100	-.212	.116	-.033	-.095	.045
Completeness of Instructional Package	.054	-.164	-.037	-.100	.098	-.103
Utilization of Student Evaluation	-.085	-.180	.027	-.061	.152	-.104
Degree of Individualization	-.063	-.197	-.038	-.040	.088	-.138
Level of Innovation	-.061	-.275	.016	-.060	.129	-.126
Number of Minutes per Day	.075	.047	-.009	-.004	-.043	-.090
r ($p < .05$) - EdExAG/Student Level[1]	.068	.065	.041	.042	.050	.049

Note: For convenience, all coefficients were truncated to three decimal places.

[1] Absolute value of the coefficient required for a correlation between one student-level variable (posttest residual) and an EdExAG-level variable to be statistically significant ($\alpha \leq .05$, two-tailed)--see equation IV-3 for the computing formula. All correlations in this table are mixed-level correlations.

significant correlations for Degree of Individualization, two were negative. Looking at the significant correlations between the individual scales of innovation and posttest residualized for pretest and SES, in all analysis samples but one (Cohort 6, SP72) the majority were negative. The Number of Minutes per Day part correlations were statistically significant in only two samples. In one sample the correlation was positive (Cohort 1, SP72) and in the other it was negative (Cohort 6, SP73).

Analysis of the Two Revised Growth Models

Two multipredictor regression models were developed to test the hypothesis that growth in arithmetic achievement was associated with the Level of Innovation or, alternatively, the Degree of Individualization, of mathematics treatment to which Project LONGSTEP students were exposed. In addition to one of these treatment attributes, the other predictor variables in each model included (1) a measure of initial achievement status in arithmetic (i.e., pretest); (2) the student's SES level; (3) the Number of Minutes per Day that his/her EdExAG group spent on arithmetic; and (4) a measure of the experience and qualifications of each student's mathematics teacher(s) (i.e., the teacher questionnaire scale, Teaching Qualifications). As with similar analyses for the language arts dependent variables, the examination of these growth models involved the utilization of regression analysis, commonality analysis and the simple correlations of the predictors with posttest.

Growth model commonality analyses. Tables V-41 through V-46 show the results of the commonality analyses of the Level of Innovation model. These analyses were based on four sets of predictors: SES (S); pretest (P); Level of Innovation and Number of Minutes per Day (E, for EdExAG variables); and Teaching Qualifications (T). Normalized coefficients, which result when a commonality coefficient is divided by the total multiple correlation squared [R^2(S,P,E,T)], were also included in these tables since they would reflect proportions of predictable variance and sum to 1.0 no matter what the magnitude of the overall R^2s in the different analysis samples. Below the overall multiple correlation squared [R^2(S,P,E,T)] for this model is the square of the multiple correlation for the regression model containing SES, pretest, the dummy variables encoding EdExAG group

TABLE [illegible]-41

Commonality Coefficients* for the Level of Innovation Growth Model -- Cohort 1, Spring 1972, Arithmetic Total Score Analysis

Commonality Coefficient	SES(S)	Pretest (P)	EdExAG Variables LI and NMIN(E)**	Teaching Qualifications(T)
$U(X_1)$	.08296 (.34112)	.10596 (.43568)	.00598 (.02457)	.00035 (.00142)
C(S,P)	.04171 (.17150)	.04171 (.17150)		
C(S,E)	-.00490 (-.02016)		-.00490 (-.02016)	
C(S,T)	.00035 (.00146			.00035 (.00146)
C(P,E)		.01178 (.04845)	.01178 (.04845)	
C(P,T)		.00062 (.00256)		.00062 (.00256)
C(E,T)			.00015 (.00060)	.00015 (.00060)
C(S,P,E)	-.00038 (-.00155)	-.00038 (-.00155)	-.00038 (-.00155)	
C(S,P,T)	-.00082 (-.00337)	-.00082 (-.00337)		-.00082 (-.00337)
C(S,E,T)	-.00032 (-.00131)		-.00032 (-.00131)	-.00032 (-.00131)
C(P,E,T)		.00074 (.00304)	.00074 (.00304)	.00074 (.00304)
C(S,P,E,T)	-.00097 (-.00401)	-.00097 (-.00401)	-.00097 (-.00401)	-.00097 (-.00401)
$R^2(X_1)$	.11763	.15864	.01207	.00010

$R^2(S,P,E,T)$ = .24319

$R^2(S,P,D,T)^+$ = .32758

---

* Below each commonality coefficient and in parentheses is the normalized coefficient; that is, the commonality coefficient divided by $R^2(S,P,E,T)$.

**LI and NMIN are abbreviations for Level of Innovation and Number of Minutes per Day.

+ The multiple correlation squared of posttest with SES, pretest, the dummy variables encoding EdExAG group membership and the three key teacher variables.

TABLE V-42

Commonality Coefficients* for the Level of Innovation Growth Model -- Cohort 1, Spring 1973, Arithmetic Total Score Analysis

Commonality Coefficient	SES(S)	Pretest (P)	EdExAG Variables LI and NMIN(E)**	Teaching Qualifications(T)
$U(X_1)$	.00012 (.00020)	.46058 (.81086)	.03693 (.06501)	.00414 (.00729)
C(S,P)	.04676 (.08233)	.04676 (.08233)		
C(S,E)	.00005 (.00009)		.00005 (.00009)	
C(S,T)	.00029 (.00052)			.00029 (.00052)
C(P,E)		-.00511 (-.00899)	-.00511 (-.00899)	
C(P,T)		.01088 (.01916)		.01088 (.01916)
C(E,T)			.00332 (.00585)	.00332 (.00585)
C(S,P,E)	-.00483 (-.00850)	-.00483 (-.00850)	-.00483 (-.00850)	
C(S,P,T)	.01141 (.02008)	.01141 (.02008)		.01141 (.02008)
C(S,E,T)	-.00046 (-.00081)		-.00046 (-.00081)	-.00046 (-.00081)
C(P,E,T)		.00260 (.00457)	.00260 (.00457)	.00260 (.00457)
C(S,P,E,T)	.00132 (.00233)	.00132 (.00233)	.00132 (.00233)	.00132 (.00233)
$R^2(X_1)$	.05467	.56362	.03383	.03351

$R^2(S,P,E,T) = .56801$

$R^2(S,P,D,T)^+ = .63544$

---

* Below each commonality coefficient and in parentheses is the normalized coefficient; that is, the commonality coefficient divided by $R^2(S,P,E,T)$.

**LI and NMIN are abbreviations for Level of Innovation and Number of Minutes per Day.

+ The multiple correlation squared of posttest with SES, pretest, the dummy variables encoding EdExAG group membership and the three key teacher variables.

TABLE V-43

Commonality Coefficients* for the Level of Innovation Growth Model -- Cohort 4, Spring 1972, Arithmetic Total Score Analysis

Commonality Coefficient	SES (S)	Pretest (P)	EdExAG Variables LI and NMIN(E)**	Teaching Qualifications (T)
$U(X_i)$	.00417 (.00657)	.53730 (.84640)	.00005 (.00008)	.00041 (.00065)
C(S,P)	.08295 (.13067)	.08295 (.13067)		
C(S,E)	.00005 (.00007)		.00005 (.00007)	
C(S,T)	.00057 (.00090)			.00057 (.00090)
C(P,E)		.00710 (.01119)	.00710 (.01119)	
C(P,T)		-.00039 (-.00061)		-.00039 (-.00061)
C(E,T)			.00008 (.00013)	.00008 (.00013)
C(S,P,E)	.00160 (.00252)	.00160 (.00252)	.00160 (.00252)	
C(S,P,T)	.00245 (.00386)	.00245 (.00386)		.00245 (.00386)
C(S,E,T)	.00011 (.00017)		.00011 (.00017)	.00011 (.00017)
C(P,E,T)		.00003 (.00005)	.00003 (.00005)	.00003 (.00005)
C(S,P,E,T)	-.00170 (-.00267)	-.00170 (-.00267)	-.00170 (-.00267)	-.00170 (-.00267)
$R^2(X_i)$	.09021	.62935	.00733	.00158

$R^2(S,P,E,T) = .63481$

$R^2(S,P,D,T)^+ = .68854$

---

* Below each commonality coefficient and in parentheses is the normalized coefficient; that is, the commonality coefficient divided by $R^2(S,P,E,T)$.

**LI and NMIN are abbreviations for Level of Innovation and Number of Minutes per Day.

+ The multiple correlation squared of posttest with SES, pretest, the dummy variables encoding EdExAG group membership and the three key teacher variables.

TABLE V-44

Commonality Coefficients* for the Level of Innovation Growth Model -- Cohort 4, Spring 1973, Arithmetic Total Score Analysis

Commonality Coefficient	SES(S)	Pretest (P)	EdExAG Variables LI and NMIN(E)**	Teaching Qualifications(T)
$U(X_i)$	.00440 (.00653)	.53237 (.79138)	.00142 (.00210)	.00020 (.00030)
C(S,P)	.08142 (.12103)	.08142 (.12103)		
C(S,E)	.00023 (.00035)		.00023 (.00035)	
C(S,T)	.00012 (.00018)			.00012 (.00018)
C(P,E)		.03790 (.05633)	.03790 (.05633)	
C(P,T)		.00081 (.00120)		.00081 (.00120)
C(E,T)			-.00012 (-.00018)	-.00012 (-.00018)
C(S,P,E)	.01197 (.01780)	.01197 (.01780)	.01197 (.01780)	
C(S,P,T)	-.00071 (-.00105)	-.00071 (-.00105)		-.00071 (-.00105)
C(S,E,T)	-.00002 (-.00002)		-.00002 (-.00002)	-.00002 (-.00002)
C(P,E,T)		.00354 (.00526)	.00354 (.00526)	.00354 (.00526)
C(S,P,E,T)	-.00081 (-.00121)	-.00081 (-.00121)	-.00081 (-.00121)	-.00081 (-.00121)
$R^2(X_i)$	.09660	.66649	.05411	.00301

$R^2(S,P,E,T) = .67272$

$R^2(S,P,D,T)^+ = .71124$

---

* Below each commonality coefficient and in parentheses is the normalized coefficient; that is, the commonality coefficient divided by $R^2(S,P,E,T)$.

**LI and NMIN are abbreviations for Level of Innovation and Number of Minutes per Day.

+ The multiple correlation squared of posttest with SES, pretest, the dummy variables encoding EdExAG group membership and the three key teacher variables.

TABLE V-45

Commonality Coefficients* for the Level of Innovation Growth Model -- Cohort 6, Spring 1972, Arithmetic Total Score Analysis

Commonality Coefficient	SES(S)	Pretest (P)	EdExAG Variables LI and NMIN(E)**	Teaching Qualifications(T)
$U(X_1)$	.00[illegible]79 (.00[illegible]62)	.56532 (.83871)	.00806 (.01195)	.00282 (.00419)
C(S,P)	.07634 (.11327)	.07634 (.11327)		
C(S,E)	-.00124 (-.00185)		-.00124 (-.00185)	
C(S,T)	-.00049 (-.00073)			-.00049 (-.00073)
C(P,E)		.01673 (.02482)	.01673 (.02482)	
C(P,T)		-.00234 (-.00348)		-.00234 (-.00348)
C(E,T)			-.00221 (-.00328)	-.00221 (-.00328)
C(S,P,E)	.00504 (.00748)	.00504 (.00748)	.00504 (.00748)	
C(S,P,T)	.00004 (.00006)	.00004 (.00006)		.00004 (.00006)
C(S,E,T)	.00055 (.00082)		.00055 (.00082)	.00055 (.00082)
C(P,E,T)		.00174 (.00259)	.00174 (.00259)	.00174 (.00259)
C(S,P,E,T)	-.00012 (-.00017)	-.00012 (-.00017)	-.00012 (-.00017)	-.00012 (-.00017)
$R^2(X_1)$	.08392	.66276	.02855	.00000

$R^2(S,P,E,T) = .67404$

$R^2(S,P,D,T)^+ = .68975$

---

* Below each commonality coefficient and in parentheses is the normalized coefficient; that is, the commonality coefficient divided by $R^2(S,P,E,T)$.

**LI and NMIN are abbreviations for Level of Innovation and Number of Minutes per Day.

+ The multiple correlation squared of posttest with SES, pretest, the dummy variables encoding EdExAG group membership and the three key teacher variables.

TABLE V-46

Commonality Coefficients* for the Level of Innovation Growth Model -- Cohort 6, Spring 1973, Arithmetic Total Score Analysis

Commonality Coefficient	SES(S)	Pretest (P)	EdExAG Variables LI and NMIN(E)**	Teaching Qualifications(T)
$U(X_1)$	.00299 (.00404)	.58436 (.78863)	.00626 (.00844)	.00006 (.00009)
C(S,P)	.07930 (.10703)	.07930 (.10703)		
C(S,E)	.00065 (.00088)		.00065 (.00088)	
C(S,T)	.00002 (.00002)			.00002 (.00002)
C(P,E)		.04855 (.06552)	.04855 (.06552)	
C(P,T)		.00052 (.00070)		.00052 (.00070)
C(E,T)			.00042 (.00056)	.00042 (.00056)
C(S,P,E)	.00414 (.00559)	.00414 (.00559)	.00414 (.00559)	
C(S,P,T)	-.00009 (-.00012)	-.00009 (-.00012)		-.00009 (-.00012)
C(S,E,T)	-.00018 (-.00025)		-.00018 (-.00025)	-.00018 (-.00025)
C(P,E,T)		.01631 (.02201)	.01631 (.02201)	.01631 (.02201)
C(S,P,E,T)	-.00234 (-.00315)	-.00234 (-.00315)	-.00234 (-.00315)	-.00234 (-.00315)
$R^2(X_1)$	.08450	.73077	.07381	.01472

$R^2(S,P,E,T)$ = .74098

$R^2(S,P,D,T)^+$ = .75048

---

* Below each commonality coefficient and in parentheses is the normalized coefficient; that is, the commonality coefficient divided by $R^2(S,P,E,T)$.

**LI and NMIN are abbreviations for Level of Innovation and Number of Minutes per Day.

+ The multiple correlation squared of posttest with SES, pretest, the dummy variables encoding EdExAG group membership and the three key teacher variables.

membership, and the three key teacher variables [$R^2(S,P,D,T)$] as predictors of posttest performance. The R^2 for this dummy variable model expresses the maximum proportion of the variation in the criterion that can be accounted for by any and all variables defined at the EdExAG group level when used along with SES, pretest and the three teacher scales as predictors of posttest performance.

Comparing the proportions of posttest arithmetic variance accounted for by the Level of Innovation model [the $R^2(S,P,E,T)$] with that yielded by the dummy variable model [$R^2(S,P,D,T)$] shows that there was a noticeable trend for the discrepancies to be inversely related to grade level. Thus, the Cohort 6 differences between $R^2(S,P,E,T)$ and $R^2(S,P,D,T)$ were considerably smaller than the differences in Cohort 4 and especially in Cohort 1. Since the same pretest and SES scores were present in both the Level of Innovation model and the dummy variable model, these differences implied that either the predictive validity of the three key teacher scales in the dummy variable model (Teacher Age, Teacher SES Background and Teaching Qualifications) was somewhat greater than that for Teaching Qualifications alone, or that the Level of Innovation superindex and Number of Minutes per Day were accounting for very little of the posttest variance among EdExAG groups. A comparison of the growth model commonality analyses with those for the dummy variable analyses shows that the proportions of posttest variance attributable to Level of Innovation and Number of Minutes per Day [the $R^2(X_1)$ in Tables V-41 through V-46], however, were considerably smaller than the proportions of variance attributable to EdExAG group membership [the $R^2(X_1)$ in Tables V-33 through V-38]. The uniquenesses followed the same pattern. Possibly because the uniquenesses for EdExAG group membership accounted for smaller proportions of variance in the older cohorts, the difference between the Level of Innovation R^2 and the dummy variable model R^2 decreased correspondingly.

Only in the Cohort 1, SP73 analysis sample was the unique association of Level of Innovation and Number of Minutes per Day notable. Although the multiple correlation squared of these two predictors with posttest was worthy of note in the Cohort 4, SP73 sample and both Cohort 6 samples, the uniquenesses in these samples were small because of the degree to which these scales were confounded with pretest.

These same associative trends were also evident in the Degree of Individualization model. Because of this similarity, these commonality analyses have been placed in Appendix V-O.

Growth model regression analyses. The regression coefficients for these two models are presented in Tables V-47 and V-48, respectively. If it is assumed that the regression coefficients index the association of a variable with growth, statistically controlling for differences with respect to initial status and the other predictors in the model, these results indicate that Level of Innovation was related to growth in arithmetic achievement but primarily in a negative manner. In the one analysis sample in which the treatment measures, Level of Innovation and Number of Minutes per Day, uniquely accounted for a meaningful proportion of posttest variance (3.6 percent in Cohort 1, SP73), the regression coefficient for Level of Innovation was statistically significant and negative, while the Number of Minutes per Day coefficient was positive. The coefficient for Teaching Qualifications was also positive and statistically significant in this sample. These results, then, would suggest that in a population of students of which the Cohort 1, SP73 sample is representative, growth in arithmetic achievement is greatest for students who are members of treatments which (1) spend relatively more time on arithmetic related activities; (2) have a relatively experienced teacher (or teachers); and (3), relative to the students in this LONGSTEP sample, are characterized by a somewhat lower Level of Innovation (as it has been defined here). Understandably enough, since the Degree of Individualization measure actually was defined by four items which were also present in the Level of Innovation superindex, examination of Table V-48 shows that these same statements may be made about the association between Degree of Individualization and growth. The high correlations between the two superindices also suggest that individualization is a major component of the innovations present in the sample, at least as measured by Level of Innovation.

Level of Innovation also received a statistically significant negative regression weight in two other analysis samples (the Cohort 4, SP73 sample

TABLE V-47

Regression Coefficients for the Level of Innovation Growth Model - Arithmetic Total Score Analyses[1]

Cohort	Posttest Year	Predictor	r	B	ß	F
1	SP72	SES	.34	1.19	.30	81.77*
		Pretest	.40	1.09	.34	104.44*
		Level of Innovation	.02	-.41	-.04	1.60
		NMIN	.10	.26	.06	2.81
		TQ	.01	-.09	-.02	0.34
		(Constant)		103.06		(df = 1; 746)
1	SP73	SES	.23	.05	.01	.20
		Pretest	.72	.82	.72	783.66*
		Level of Innovation	-.12	-2.48	-.20	61.43*
		NMIN	.12	.28	.05	4.54*
		TQ	.18	.46	.07	7.05*
		(Constant)		103.70		(df = 1; 735)
4	SP72	SES	.30	.39	.07	22.14*
		Pretest	.79	1.16	.77	2849.87*
		Level of Innovation	-.06	.08	.01	.13
		NMIN	-.06	-.04	-.01	.16
		TQ	-.04	-.15	-.02	2.20
		(Constant)		-44.72		(df = 1; 1937)
4	SP73	SES	.31	.43	.07	26.30*
		Pretest	.82	.86	.79	3184.99*
		Level of Innovation	-.23	-.63	-.04	8.45*
		NMIN	-.01	.02	.00	.09
		TQ	.05	-.11	-.01	1.21
		(Constant)		70.78		(df = 1; 1958)

(continued)

TABLE V-47 (continued)

Cohort	Posttest Year	Predictor	r	B	β	F
6	SP72	SES	.29	.49	.07	16.70*
		Pretest	.81	1.00	.79	2492.20*
		Level of Innovation	.10	4.53	.10	33.74*
		NMIN	-.15	.06	.01	.18
		TQ	.00	.63	.06	12.44*
		(Constant)		-194.86		(df = 1; 1437)
6	SP73	SES	.29	.46	.06	17.33*
		Pretest	.85	.91	.82	3381.85*
		Level of Innovation	-.22	-1.42	-.07	23.22*
		NMIN	-.16	-.37	-.05	12.16*
		TQ	.12	-.08	-.01	.37
		(Constant)		79.84		(df = 1; 1499)

[1] Notation and abbreviation:

r = zero-order correlation with posttest

B = raw score regression coefficient

β = standard score regression coefficient

F = conventional F test of the regression coefficient

NMIN = EdExAC scale, Number of Minutes per Day (spent on mathematics)

TQ = teacher scale, Teaching Qualifications

*p ≤ .05

TABLE V-48

Regression Coefficients for the Degree of Individualization Growth Model - Arithmetic Total Score Analyses[1]

Cohort	Posttest Year	Predictor	r	B	β	F
1	SP72	SES	.34	1.17	.30	81.64*
		Pretest	.40	1.08	.34	104.34*
		Degree of Individualization	-.01	-.97	-.04	1.59
		NMIN	.10	.27	.06	3.05
		TQ	.01	-.06	-.01	.16
		(Constant)		100.48		(df = 1; 746)
1	SP73	SES	.23	-.00	-.00	.00
		Pretest	.72	.82	.72	750.33*
		Degree of Individualization	-.07	-4.37	-.15	32.50*
		NMIN	.12	.36	.07	7.02*
		TQ	.18	.46	.07	6.64*
		(Constant)		89.18		(df = 1; 735)
4	SP72	SES	.30	.39	.07	22.10*
		Pretest	.79	1.15	.77	2866.66*
		Degree of Individualization	-.06	-.95	-.03	3.91*
		NMIN	-.06	-.04	-.01	.15
		TQ	-.04	-.19	-.03	3.60
		(Constant)		-30.88		(df = 1; 1937)
4	SP73	SES	.31	.44	.07	26.60*
		Pretest	.82	.87	.79	3272.26*
		Degree of Individualization	-.18	-.95	-.03	3.93*
		NMIN	-.01	.01	.00	.01
		TQ	.05	-.10	-.01	1.09
		(Constant)		63.75		(df = 1; 1958)

(continued)

TABLE V-48 (continued)

Cohort	Posttest Year	Predictor	r	B	β	F
6	SP72	SES	.29	.48	.06	15.33*
		Pretest	.81	1.00	.79	2453.48*
		Degree of Individualization	.03	5.07	.06	14.72*
		NMIN	-.15	-.19	-.02	2.01
		TQ	.00	.30	.03	3.35
		(Constant)		-89.39		(df = 1; 1437)
6	SP73	SES	.29	.47	.06	18.01*
		Pretest	.85	.91	.82	3397.42*
		Degree of Individualization	-.23	-3.12	-.08	27.70*
		NMIN	-.16	-.36	-.05	11.41*
		TQ	.12	-.13	-.01	.90
		(Constant)		79.44		(df = 1; 1499)

[1] Notation and abbreviation:

r = zero-order correlation with posttest
B = raw score regression coefficient
β = standard score regression coefficient
F = conventional F test of the regression coefficient
NMIN = EdExAG scale, Number of Minutes per Day (spent on mathematics)
TQ = teacher scale, Teaching Qualifications

$^{*}p \leq .05$

and the Cohort 6, SP73 sample). In this latter data set, Number of Minutes per Day also received a significant negative coefficient. These same results occurred in the Degree of Individualization model, where an additional sample (Cohort 4, SP72) also yielded a negative regression coefficient for Degree of Individualization.

It would be misleading, however, to state that these regression models were unambiguously demonstrating the finding that less innovation was related to greater amounts of arithmetic achievement. Level of Innovation or Degree of Individualization were both positively related to growth in the Cohort 6, SP72 analyses. (The significant regression coefficient for Teaching Qualifications in the Level of Innovation analysis for this sample is difficult to interpret because this variable's overall relationship with posttest was nil. Presumably, in this context, the Teaching Qualifications variable is functioning as a suppressor.)

When these results are compared to those involving adjustment of posttest for pretest and SES differences (i.e., the posttest residual), it is clear that roughly the same results are obtained whether or not the criterion or the predictors are residualized. Since the majority of these analyses showed that Level of Innovation or Degree of Individualization (as defined here) was negatively, rather than positively, related to growth (if at all), it would seem that this is the most notable trend in these samples as a whole. The results for Number of Minutes per Day and Teaching Qualifications did show some tendency to be positively related to student achievement in arithmetic. However, the hypothesis that a higher Level of Innovation or a greater Degree of Individualization (as defined in this study) is positively and consistently related to student achievement in arithmetic is clearly not supported by our data.

Examination of Unusually Effective EdExAG Groups

For purposes of this exploratory analysis, the EdExAG groups with the greatest positive and the greatest negative mean residual scores in each year were selected for examination, after eliminating those groups with very few students. Two independent analyses were performed for each cohort, the Spring 1972 posttest analysis and the Spring 1973 posttest analysis.

This section discusses the outlier group analyses for all three cohorts. The tables for Cohort 1 are presented here for purposes of illustration. The corresponding tables for Cohorts 4 and 6 are contained in Appendix V-P.

The key analysis variables for the Cohort 1 positive and negative outlier groups for the Spring 1972 Arithmetic Total analysis are presented in Table V-49. All outlier groups, the three positive and the two negative, were above the norm on pretest; but only the positive outlier groups (the high performing ExAGs) exceeded the norm on posttest. Using the grade equivalent score for the average student in each group as a crude indicator of the difference between the groups on posttest, the positive outlier groups (grade 2) ranged between 3.1 and 3.3; the two negative outliers ranged between 2.2 and 2.5. Looking at student characteristics, we find generally that the positive outlier groups were slightly higher in average SES level and contained a much higher proportion of white students. A slightly lower proportion of girls was present in the positive outlier groups as compared with the negative outliers.

In the Spring 1973 analysis, the students in two of the three positive outliers averaged above the norm on pretest, and all three positive outlier groups had mean scores above the norm on posttest. For the negative outliers, two of the three groups averaged above the norm on both pretest and posttest. The difference between the two groups is indicated roughly by the average yearly growth which, for students in the positive outlier groups, ranged from 1.3 to 2.0 grade equivalents compared with yearly gains of from 0.9 to 1.0 for the negative outlier groups.

In terms of SES, sex and ethnic group membership, the positive and negative outlier groups in 1973 were roughly comparable, with no consistent differences. Considering student attitudes, the students in the positive outliers expressed a less favorable Attitude toward Math than did students in the negative outlier groups during Spring 1973, the first time that attitude measures were obtained from students in this cohort.

Comparing Tables V-49 and V-50, we find no evidence of either treatment variables or teacher variables which consistently and dramatically differentiated between the high performing groups and the lower performing

TABLE V-49

Comparison of Key Analysis Variables for Positive and Negative Outlier EdExAG Groups-- Cohort 1, Spring 1972, Arithmetic Total Score Analysis

Variable		Greatest Negative Outliers		Greatest Positive Outliers		
Number of Students		19	13	72	62	103
Posttest Residualized for Pretest and SES	$\bar{X}$	-24.70	-29.16	14.97	14.64	7.15
	SD	30.81	22.66	34.63	32.85	33.73
	g^1	0.39	0.30	0.08	0.21	0.86
Pretest (National Norm MA=85)	$\bar{X}$	92.22	88.79	96.52	91.20	90.78
	SD	10.93	12.44	10.89	13.79	9.70
	g	0.02	0.84	0.18	0.13	-0.08
Posttest (National Norm=303)	$\bar{X}$	290.10	275.02	337.49	327.51	328.69
	SD	36.95	20.17	36.24	39.83	38.46
	g	0.85	0.04	0.20	0.23	0.78
SES	$\bar{X}$	96.73	90.66	99.34	96.03	104.25
	SD	7.64	5.51	9.32	7.57	9.14
	g	1.08	-1.44	0.09	0.82	-0.42
Ethnic Group[2] (white=1;non-white=0)	$\bar{X}$	0.53	0.69	0.99	0.97	0.87
Sex[2] (girl=1;boy=0)	$\bar{X}$	0.63	0.62	0.54	0.53	0.59
Utilization of Objectives		3.00	2.33	3.00	1.00	3.00
Individualization in Decision Making		1.00	1.00	1.00	1.00	1.67

---

[1] Index of skewness

$$g = \frac{\frac{\Sigma x^3}{N}}{\frac{\Sigma x^2}{N}\sqrt{\frac{\Sigma x^2}{N}}}$$

, where g = skewness index,

x = deviation of a score from $\bar{X}$, where $\bar{X}$ is an EdExAG group mean for variable X, and

N = number of students.

[2] For binary coded variables the $\bar{X}$ is equal to the proportion of students coded with a "1."

TABLE V-49 (continued)

Variable	Greatest Negative Outliers		Greatest Positive Outliers		
Teacher or Locally Developed Materials	1.00	2.00	1.50	1.50	2.00
Individualization of Instructional Pace	2.00	1.00	3.00	1.00	2.00
Scheduling Characteristics	1.00	1.00	3.00	1.00	2.00
Use of Performance Agreements	1.00	1.00	2.00	1.00	2.00
Classroom Group Organization	2.00	2.00	2.00	1.00	2.00
Teaching Unit Composition	2.50	2.00	1.50	1.00	3.00
Completeness of Instructional Package	2.00	1.50	3.00	2.00	2.50
Utilization of Student Evaluation	2.00	2.67	2.00	1.00	3.00
Level of Innovation	17.50	16.50	22.00	11.50	23.17
Degree of Individualization	6.00	5.67	8.00	4.00	8.67
Number of School Days per Year	180.00	180.00	180.00	180.00	200.00
Number of Minutes per Day	90.00	105.00	90.00	60.00	90.00
Number of Teachers $\bar{X}$	4.00	1.00	1.00	1.00	6.00
Teacher Age $\bar{X}$	29.50	24.00	25.33	30.00	29.67
Teacher Socioeconomic Background $\bar{X}$	110.35	117.86	108.32	106.60	95.21
Teaching Qualifications $\bar{X}$	89.91	86.49	87.69	100.59	98.76

TABLE V-50

Comparison of Key Analysis Variables for Positive and Negative Outlier EdExAG Groups-- Cohort 1, Spring 1973, Arithmetic Total Score Analysis

Variable		Greatest Negative Outliers			Greatest Positive Outliers		
Number of Students		60	9	112	39	29	51
Posttest Residualized for Pretest and SES	$\bar{X}$	-8.45	-12.08	-15.06	34.36	28.69	6.71
	SD	21.89	29.28	31.98	24.20	25.20	23.43
	g^1	-0.28	-0.82	-0.13	-0.38	-0.34	0.71
Pretest (National Norm=303)	$\bar{X}$	312.80	287.40	327.44	312.91	297.60	314.54
	SD	33.26	21.47	36.45	32.27	32.88	26.40
	g	0.69	-0.35	0.70	-0.34	0.60	0.21
Posttest (National Norm=358)	$\bar{X}$	364.17	339.68	369.52	407.08	388.82	380.74
	SD	34.96	36.76	47.93	31.32	40.81	32.76
	g	0.12	0.97	-0.02	-0.05	-0.41	0.28
Attitude toward Math, SP73	$\bar{X}$	97.50		101.64	93.27		100.04
	SD	12.13		10.31	11.63		10.70
	g	0.01		-0.54	0.64		-0.28
SES	$\bar{X}$	93.42	98.10	104.06	93.38	95.03	96.28
	SD	4.74	10.15	9.32	5.92	7.26	7.70
	g	1.45	0.48	-0.45	0.81	0.99	-0.04
Ethnic Group[2] (white=1; non-white=0)	$\bar{X}$	0.93	0.67	0.86	0.95	0.42	0.76
Sex[2] (girl=1;boy=0)	$\bar{X}$	0.53	0.44	0.61	0.31	0.65	0.51

(continued)

Note: Blank entries indicate no data collected.

[1] Index of skewness

$$g = \frac{\frac{\Sigma x^3}{N}}{\frac{\Sigma x^2}{N}\sqrt{\frac{\Sigma x^2}{N}}}$$

, where g = skewness index,

x = deviation of a score from $\bar{X}$, where $\bar{X}$ is an EdExAG group mean for variable X, and

N = number of students.

[2] For binary coded variables the $\bar{X}$ is equal to the proportion of students coded with a "1."

TABLE V-50 (continued)

Variable	Greatest Negative Outliers			Greatest Positive Outliers		
Utilization of Objectives	3.00	1.67	3.00	1.00	1.00	3.00
Individualization in Decision Making	2.00	1.67	1.67	1.00	1.00	1.00
Teacher or Locally Developed Materials	2.00	2.50	2.50	2.00	2.00	1.00
Individualization of Instructional Pace	2.00	2.00	2.00	2.00	2.00	2.00
Scheduling Characteristics	3.00	2.00	2.00	1.00	1.00	1.00
Use of Performance Agreements	3.00	1.00	2.00	1.00	1.00	1.00
Classroom Group Organization	3.00	2.00	2.00	2.00	2.00	3.00
Teaching Unit Composition	2.50	1.00	3.00	1.00	1.00	2.00
Completeness of Instructional Package	2.00	2.00	2.50	1.50	1.50	1.00
Utilization of Student Evaluation	2.67	2.67	3.00	2.33	1.00	2.33
Level of Innovation	25.17	18.50	23.67	14.83	13.50	17.33
Degree of Individualization	9.67	7.34	8.67	6.33	5.00	6.33
Number of School Days per Year	180.00	180.00	200.00	180.00	180.00	180.00
Number of Minutes per Day	40.00	45.00	45.00	60.00	30.00	45.00
Teacher/Student Contact Ratio[3]	3.30	4.44	6.67	3.45	3.85	3.13

(continued)

---

[3] Relative measure of teacher/student contact equal to the percentage of teacher time per individual student.

TABLE V-50 (continued)

Variable		Greatest Negative Outliers			Greatest Positive Outliers		
Number of Teachers	$\bar{X}$	1.00	1.00	1.00	1.00	1.00	1.04
Teacher Age	$\bar{X}$	43.97	26.67	34.36	30.00	60.00	26.75
Teacher Socioeconomic Background	$\bar{X}$	100.36	115.60	97.07	91.29	108.04	103.52
Teaching Qualifications	$\bar{X}$	96.98	87.68	101.31	100.40	109.96	91.25

groups over both years. While two of the three positive outliers in 1972 exhibited a higher Level of Innovation and a greater Degree of Individualization than did the negative outliers, in 1973 the positive outliers were all lower on both superscales. Teachers of the high performing EdExAGs in 1972 tended to score higher on the Teaching Qualifications scale and lower on the Teacher Socioeconomic Background scale; however, in 1973 the teachers of both groups displayed no consistent differences in Teaching Qualifications, Age or Socioeconomic Background.

In looking over the other profile variables listed in Appendix IV-H, we again found no scales which showed the same pattern of differentiation between positive and negative outliers for both years.

For the Cohort 4 outlier groups, the key analysis variables for the Spring 1972 Arithmetic Total Score analysis are in Appendix V-P, Table V-P-1. From the standpoint of student performance on the arithmetic test, only one of the positive outliers averaged noticeably above the norm on pretest, as compared with two of the negative outliers. On posttest, however, all three positive outlier groups were above the norm; the negative outliers all averaged below the norm. The average yearly gains for students in the positive outlier groups ranged between 1.6 and 1.9 grade equivalents, in contrast with average yearly gains of 0.3 and 0.5 grade equivalents for the low performers. Looking at student characteristics, all the outliers had very high percentages of white students except for one negative outlier, which still had 60 percent whites. Both the SES level and the sex distribution in the positive and negative outliers were mixed. The groups differed little with respect to Attitude toward Math.

Of the three positive outliers identified in 1973, two averaged below the national norm on pretest (see Appendix V-P, Table V-P-2). All three were above the norm on posttest. In contrast, all three negative outliers averaged above the norm on pretest but only one was above on posttest. Some indication of the differences in yearly growth between the groups is given by the fact that the gains of the average student in the positive outlier groups ranged between 1.7 and 1.8 grade equivalents, whereas gains in the negative outlier groups ranged from 0.1 to 0.5. The outlier groups also appeared to differ in terms of SES level, with two of three positive outlier groups noticeably below the SES levels of the negative outliers. However, the groups differed little on either proportion of minority students or of girls. (The lowest proportion of whites and the highest proportion of girls, though, were present in one positive outlier group.) Looking at Attitude toward Math, it appears that students in the positive outlier groups were more often lower in this scale at the end of both fifth grade and sixth grade than were students in the lower performing groups.

Summarizing these outlier analyses for Cohort 4, there seems to be no clear indication of consistent differences in either year. Similarly, in terms of teacher characteristics, none appeared to distinguish the positive from the negative outliers consistently in both years. This was also true for the other profile variables listed in Appendix IV-H.

For Cohort 6, two negative outlier groups and three positive outlier groups were selected for analysis in 1972 (see Appendix V-P, Table V-P-3). Only one group, a positive outlier, had a mean score below the national norm on pretest. This same group was also below the norm on posttest, as were both of the negative outlier groups. The contrast between the positive and negative outlier groups is indicated by their average yearly gains: for the positive outliers, the gains reflected about 1.0 grade equivalent of growth; for the negative outliers, the yearly gains were 0.2 or less of a grade equivalent.

The students in the positive outliers were more often higher on the SES scale; however, the average SES level of one positive outlier was well below all other groups. In terms of ethnic composition, both negative outlier groups contained approximately 80 percent whites; one positive outlier had less than 40 percent white students and another was composed almost entirely of whites. No major differences were exhibited between the positive and negative groups in the proportion of girls. In 1971, all positive outlier groups averaged higher than the negative groups in terms of Attitude toward Math. It was not possible to compare the positive outlier groups with the negative groups in 1972 since attitude data were not available from the two negative outliers in that year.

The key analysis variables for the 1973 outlier group comparisons are in Appendix V-P, Table V-P-4. In the Spring 1973 analysis, two of the three negative outliers and one of the three positive outlier groups averaged below the national norm on both pretest and posttest. The average yearly gains in Arithmetic Total Score for the positive outlier groups ranged between 0.7 and 1.2 in grade equivalent units, whereas the gains for the negative outliers ranged between 0.1 and 0.4 grade equivalents.

Looking at the characteristics of the students in these groups we find that the average SES levels were more often higher in the positive outliers, as was the percentage of whites. Both the positive and negative outliers typically had over 70 percent white students, but the greatest negative outlier had only 39 percent white students. The positive outliers more often had a slightly higher proportion of girls than did the negative outlier groups. On the whole, no major differences were exhibited on the Attitude toward Math scale.

Considering differences in treatment variables between the positive and negative outlier groups, the negative outliers in 1972 had both the highest and the lowest Level of Innovation scale scores. The results for the Degree of Individualization scale were similarly ambiguous. In 1973, however, two of the three positive outlier groups were very definitely lower on both Level of Innovation and Degree of Individualization; the third closely matched the scale values exhibited by all three negative outliers. Fairly mixed results also appeared with respect to the individual scales of innovation.

Considering the teacher characteristics of Age, Socioeconomic Background, and Teaching Qualifications, none appeared to differentiate clearly between the positive and the negative outlier groups in both years. The only such variable that showed the least tendency for consistency was that of Teaching Qualifications. In both years, the positive outliers, on the whole, tended to have a slightly higher score in this scale. Several of the other profile variables, as listed in Appendix IV-H, differentiated between the positive and negative outliers in a single year, but none did so consistently over both analysis periods.

Reviewing the six separate analyses of outlier groups with respect to the Arithmetic Total Score (Spring 1972 and Spring 1973 analyses for each of three cohorts) provides some evidence of the degree of consistency present in our outlier data. In terms of the overall Level of Innovation, four of the six analyses showed that the positive outlier groups typically scored lower. Only in the Spring 1972 analyses for Cohort 1 were the better performing groups more often characterized by a higher Level of Innovation. This same pattern was present for the superscale, Degree of Individualization.

The results for teacher characteristics were not nearly so definite. Of the scales, Teaching Qualifications, Teacher Socioeconomic Background and Teacher Age, only Teaching Qualifications gave some hint of discriminating between the positive and the negative outliers over the two years. This occurred only in Cohort 6, where the positive outliers, on the whole, tended to have teachers with slightly higher Teaching Qualifications.

Examination of Unusual Student Growth Patterns in Mathematics Achievement

The students in Cohorts 1, 4 and 6 who completed the arithmetic sections of their achievement tests for all three years of Project LONGSTEP were subdivided according to their deviations from the mean residualized criterion scores in the two analysis years, 1972 and 1973. Students whose residual scores were equal to or more than one-half of a standard deviation below the mean of their cohort's residualized CTBS Arithmetic Test Total Expanded Scale Scores for both analysis years were termed "consistent underachievers in mathematics." Students whose residual posttest scores were equal to or more than one-half of a standard deviation above the mean on the residualized Arithmetic Total Score for both analysis years were termed "consistent overachievers in mathematics." The following discussion relates to these two groups. Students who exhibited unusual but inconsistent patterns (over- then underachievement and under- then overachievement) will be mentioned later.

Table V-51 shows the average characteristics of the Cohort 1 consistent overachievers and underachievers in mathematics. A comparison between the two groups on student background indicates that there were almost no differences between them with respect to SES level, proportion of girls and ethnic composition. A slight difference between the underachievers and overachievers with respect to Attitude toward Math was present in Spring 1973. Since this variable was first measured at the time of the final achievement criterion, earlier attitudinal measures were not available for this cohort.

The students in Cohort 1 who were consistent underachievers in mathematics had an average CTMM Mental Age - Non-Language of about 90 months at the end of first grade (Cohort 1, SP72), and average CTBS Arithmetic Total Expanded Scale Scores of 283 (grade equivalent 2.3) at the end of the second grade and 315 (grade equivalent 2.9) at the end of third grade. The Cohort 1 consistent overachievers in mathematics had an average Mental Age - Non-Language of 94 months at the end of first grade, and average Arithmetic Total scores of 353 (grade equivalent 3.6) at the end of the second grade and 436 (grade equivalent 5.6) at the end of the third grade. Thus, by the end of the third grade, the Cohort 1 overachievers in mathematics were approximately 2.7 grade equivalents above the underachievers.

An examination of the overall Level of Innovation to which these two groups were exposed during the analysis years indicates that, during the second grade, the Cohort 1 consistent overachievers in mathematics were exposed to, on the average, only a very slightly lower degree of innovative educational practices than the underachievers. The two groups were also very similar as to average Degree of Individualization during these years.

TABLE V-51

Comparison of Key Analysis Variables for Students Exhibiting Unusual Longitudinal Growth Patterns In Two Consecutive School Years--Cohort 1, Arithmetic Total Score Analysis

Variable		Consistent Underachievers		Consistent Overachievers	
Number of Students		49		47	
SES	$\bar{X}$	101.32		99.49	
	SD	9.92		9.66	
	g^1	-0.14		0.10	
Sex[2] (girl=1; boy=0)	$\bar{X}$	0.57		0.55	
Ethnic Group[2] (white=1; non-white=0)	$\bar{X}$	0.84		0.85	
		Analysis Year		Analysis Year	
		SP72	SP73	SP72	SP73
Pretest (Nat'l Norm MA=85 for SP72; Nat'l Norm=303 for SP73)	$\bar{X}$	89.50	283.47	93.90	352.60
	SD	12.30	17.42	10.49	20.64
	g	0.06	0.23	0.09	-0.08
Posttest(Nat'l Norm =303 for SP72; Nat'l Norm=358 for SP73)	$\bar{X}$	283.47	314.55	352.60	436.12
	SD	17.42	25.71	20.64	21.56
	g	0.23	0.01	-0.08	-0.02
Posttest Residualized for Pretest and SES	$\bar{X}$	-33.36	-34.43	32.85	30.94
	SD	13.17	19.60	11.88	13.74
	g	-0.66	-1.50	1.02	1.08
Attitude toward Math at Time of Pretest	$\bar{X}$				
	SD				
	g				

(continued)

Note: Blank entries indicate no data collected.

[1] Index of skewness

$$g = \frac{\frac{\Sigma x^3}{N}}{\frac{\Sigma x^2}{N}\sqrt{\frac{\Sigma x^2}{N}}}$$

, where g = skewness index,

x = deviation of a score from $\bar{X}$, where $\bar{X}$ is that group's mean on variable X, and

N = number of students.

[2] For binary coded variables the $\bar{X}$ is equal to the proportion of students coded with a "1."

TABLE V-51 (continued)

Variable		Consistent Underachievers SP72	Consistent Underachievers SP73	Consistent Overachievers SP72	Consistent Overachievers SP73
Attitude toward Math at Time of Posttest	X̄		98.10		101.64
	SD		11.40		10.23
	g		-0.10		-0.63
Utilization of Objectives	X̄	2.81	2.85	2.46	2.46
	SD	0.56	0.45	0.88	0.81
Individualization in Decision Making	X̄	1.46	1.61	1.46	1.43
	SD	0.62	0.57	0.52	0.49
Teacher or Locally Developed Materials	X̄	1.59	1.77	1.68	1.99
	SD	0.46	0.53	0.51	0.36
Individualization of Instructional Pace	X̄	2.31	2.37	2.23	2.17
	SD	0.68	0.48	0.59	0.38
Scheduling Characteristics	X̄	1.65	1.94	1.62	1.64
	SD	0.72	0.68	0.60	0.67
Use of Performance Agreements	X̄	1.55	1.84	1.60	1.87
	SD	0.57	0.79	0.57	0.79
Classroom Group Organization	X̄	2.12	2.45	2.04	2.17
	SD	0.77	0.50	0.65	0.38
Teaching Unit Composition	X̄	2.07	2.19	2.01	2.22
	SD	0.65	0.62	0.82	0.76
Completeness of Instructional Package	X̄	2.29	2.26	2.13	2.18
	SD	0.61	0.59	0.66	0.51
Utilization of Student Evaluation	X̄	2.42	2.48	2.46	2.61
	SD	0.60	0.51	0.66	0.41
Level of Innovation	X̄	20.28	21.74	19.69	20.75
	SD	3.76	2.83	4.86	3.74

(continued)

TABLE V-51 (continued)

Variable		Consistent Underachievers SP72	Consistent Underachievers SP73	Consistent Overachievers SP72	Consistent Overachievers SP73
Degree of Individualization	$\bar{X}$	7.74	8.30	7.75	8.08
Number of School Days per Year	$\bar{X}$	184.65	184.65	186.38	186.38
	SD	9.33	9.33	9.86	9.86
Number of Minutes per Day	$\bar{X}$	50.64	50.20	51.11	51.17
	SD	8.36	7.35	7.96	7.66
Teacher/Student Contact Ratio[3]	$\bar{X}$		4.24		4.47
	SD		1.48		1.59
Number of Teachers	$\bar{X}$	2.53	1.06	2.34	1.17
	SD	2.04	0.24	2.01	0.43
Teacher Age	$\bar{X}$	30.79	36.83	32.13	36.45
	SD	5.67	12.72	8.76	11.20
Teacher Socioeconomic Background	$\bar{X}$	102.39	103.35	100.70	98.97
	SD	7.64	10.72	7.42	11.10
Teaching Qualifications	$\bar{X}$	96.66	101.36	98.66	102.47
	SD	8.19	6.45	7.15	5.07

[3] Relative measure of teacher/student contact equal to the percentage of teacher time per individual student

As shown in Table V-51, the two groups were very close in terms of the length of their school year. Similarly, there were no notable differences between the groups with respect to the average amount of time per day they were exposed to mathematics instruction during the second or third grade.

The two groups were generally similar to each other with respect to teacher variables. In the second grade, the Cohort 1 overachievers tended to have slightly older, more qualified teachers; but these differences were less notable in the third grade. Teachers of the overachievers, on the average, tended to be from a lower SES background than teachers of underachievers. The two groups were similar in terms of number of teachers involved in mathematics instruction and teacher/student contact.

Table V-52 summarizes the attributes of the Cohort 4 consistent underachievers and overachievers in mathematics. The two groups were fairly close in terms of SES level and proportion of girls, but the overachievers had a higher proportion of whites (96 percent) than the underachievers (87 percent). Consistent overachievers had a more favorable Attitude toward Math than the underachievers at both pretest and posttest time.

The consistent underachievers in Cohort 4 had an average CTBS Arithmetic Test Total Expanded Scale Score at the end of the fourth grade (Cohort 4, SP72) of 397 (grade equivalent 4.6), at the end of the fifth grade of 402 (grade equivalent 4.7), and at the end of the sixth grade of 404 (grade equivalent 4.8). Thus, they could be characterized as showing no major improvement in mathematics skills over the two analysis years. In comparison, the consistent overachievers had an average arithmetic score at the end of the fourth grade of 416 (grade equivalent 5.0), at the end of the fifth grade of 493 (grade equivalent 7.2), and at the end of the sixth grade of 555 (grade equivalent 8.9). By the end of the sixth grade, then, these two groups differed from each other in average mathematics achievement by 4.1 grade equivalents.

There was no major difference between the two groups with respect to Level of Innovation or Degree of Individualization in either the fifth or

TABLE V-52

Comparison of Key Analysis Variables for Students Exhibiting Unusual Longitudinal Growth Patterns In Two Consecutive School Years--Cohort 4, Arithmetic Total Score Analysis

Variable		Consistent Underachievers		Consistent Overachievers	
Number of Students		87		90	
SES	$\bar{X}$	99.27		98.50	
	SD	9.92		9.57	
	g^1	0.33		0.30	
Sex^2 (girl=1; boy=0)	$\bar{X}$	0.46		0.49	
Ethnic Group^2 (white=1; non-white=0)	$\bar{X}$	0.87		0.96	
		Analysis Year		Analysis Year	
		SP72	SP73	SP72	SP73
Pretest (Nat'l Norm =400 for SP72; Nat'l Norm=440 for SP73)	$\bar{X}$	397.14	402.35	416.06	493.28
	SD	29.25	38.92	28.89	36.61
	g	0.40	0.56	-0.71	-0.55
Posttest (Nat'l Norm=440 for SP72; Nat'l Norm=476 for SP73)	$\bar{X}$	402.35	404.20	493.28	554.96
	SD	38.22	39.50	36.61	34.81
	g	0.56	0.66	-0.55	-0.44
Posttest Residualized for Pretest and SES	$\bar{X}$	-35.58	-34.37	33.80	34.49
	SD	15.78	15.11	17.99	17.98
	g	-1.12	-0.90	1.48	1.96
Attitude toward Math at Time of Pretest	$\bar{X}$	97.46	95.86	100.49	100.35
	SD	10.91	9.97	9.70	10.11
	g	-0.16	0.16	-0.46	-0.42

(continued)

Note: Blank entries indicate no data collected.

[1] Index of skewness

$$g = \frac{\frac{\Sigma x^3}{N}}{\frac{\Sigma x^2}{N}\sqrt{\frac{\Sigma x^2}{N}}}$$

, where g = skewness index,

x = deviation of a score from $\bar{X}$, where $\bar{X}$ is that group's mean on variable X, and

N = number of students.

[2] For binary coded variables the $\bar{X}$ is equal to the proportion of students coded with a "1."

TABLE V-52 (continued)

Variable		Consistent Underachievers SP72	Consistent Underachievers SP73	Consistent Overachievers SP72	Consistent Overachievers SP73
Attitude toward Math at Time of Posttest	$\bar{X}$	95.86	99.03	100.35	102.70
	SD	9.97	10.36	10.11	10.26
	g	0.16	0.09	-0.42	-0.45
Utilization of Objectives	$\bar{X}$	2.49	2.61	2.29	2.43
	SD	0.82	0.72	0.84	0.74
Individualization in Decision Making	$\bar{X}$	1.25	1.23	1.31	1.13
	SD	0.32	0.40	0.35	0.29
Teacher or Locally Developed Materials	$\bar{X}$	1.77	1.92	1.74	1.81
	SD	0.51	0.64	0.54	0.53
Individualization of Instructional Package	$\bar{X}$	2.02	2.34	2.02	2.10
	SD	0.69	0.76	0.60	0.70
Scheduling Characteristics	$\bar{X}$	1.45	1.45	1.50	1.48
	SD	0.54	0.67	0.60	0.54
Use of Performance Agreements	$\bar{X}$	1.47	1.72	1.60	1.82
	SD	0.66	0.85	0.66	0.78
Classroom Group Organization	$\bar{X}$	1.85	1.87	2.03	1.93
	SD	0.62	0.72	0.69	0.65
Teaching Unit Composition	$\bar{X}$	1.51	1.87	1.81	1.78
	SD	0.60	0.63	0.72	0.61
Completeness of Instructional Package	$\bar{X}$	2.07	2.30	1.90	1.92
	SD	0.61	0.57	0.60	0.55
Utilization of Student Evaluation	$\bar{X}$	2.47	2.52	2.36	2.52
	SD	0.63	0.54	0.53	0.38
Level of Innovation	$\bar{X}$	18.35	19.84	18.56	18.93
	SD	3.70	4.09	3.43	3.37

(continued)

TABLE V-52 (continued)

Variable		Consistent Underachievers		Consistent Overachievers	
		SP72	SP73	SP72	SP73
Degree of Individualization	$\bar{X}$	7.21	7.81	7.29	7.57
Number of School Days per Year	$\bar{X}$	180.97	181.06	181.42	181.20
	SD	5.82	5.89	5.44	5.65
Number of Minutes per Day	$\bar{X}$	52.45	50.46	52.41	52.72
	SD	7.89	7.86	7.81	13.02
Teacher/Student Contact Ratio[3]	$\bar{X}$		3.78		3.67
	SD		0.63		0.52
Number of Teachers	$\bar{X}$	1.30	1.17	1.67	1.51
	SD	1.03	0.94	1.17	1.11
Teacher Age	$\bar{X}$	37.30	38.90	37.10	39.29
	SD	10.83	12.45	9.54	10.35
Teacher Socioeconomic Background	$\bar{X}$	99.04	95.91	99.21	99.88
	SD	8.80	9.29	9.64	8.86
Teaching Qualifications	$\bar{X}$	103.95	99.79	103.27	101.10
	SD	7.10	9.36	6.81	8.81

[3] Relative measure of teacher/student contact equal to the percentage of teacher time per individual student

the sixth grades. During the sixth grade, however, the overachievers were exposed to a slightly lower Level of Innovation than the underachievers.

The two groups were very close in terms of amount of mathematics instruction; on the average, they had about the same length of school year and Number of Minutes per Day spent on mathematics instruction. Also, the two groups appear to have had teachers who were similar with respect to Teacher Age and Teaching Qualifications. Teacher/Student Contact Ratio differences among the groups were small.

Table V-53 shows the results for the students in Cohort 6 with consistent unusual patterns of achievement in mathematics. The consistent overachievers were below the underachievers in average SES, were more apt to be white (94 percent) than the underachievers (79 percent), and had more girls among them (61 percent) than did the underachievers (52 percent). The mean Attitude toward Math of the overachieving group was above that for the underachieving group both at pretest and at posttest time.

The Cohort 6 consistent underachievers in mathematics averaged 493 on the CTBS Arithmetic Total Test Expanded Scale Score (grade equivalent 7.2) at the end of the sixth grade (Cohort 6, SP72), 470 at the end of the seventh grade (grade equivalent 6.5), and 451 at the end of the eighth grade (grade equivalent 6.0). Relative to CTBS national norms, this represented a loss in mathematics achievement of 1.2 grade equivalents over a two-year period. In contrast, the students in Cohort 6 who were consistent overachievers during our two analysis years went from an average arithmetic scale score of 492 at the end of the sixth grade (grade equivalent 7.1) to an average scale score of 554 at the end of the seventh grade (grade equivalent 8.9), and then to an average scale score of 610 at the end of the eighth grade (grade equivalent 10.5). Thus, as the consistent underachievers were falling below grade level, the consistent overachievers were rising far above it. By the end of the eighth grade, the mean difference in mathematics achievement between these two groups corresponded to about 4.5 grade equivalents.

There was no difference between the two groups on Level of Innovation or Degree of Individualization for the first analysis year, but in the

TABLE V-53

Comparison of Key Analysis Variables for Students Exhibiting Unusual Longitudinal Growth Patterns In Two Consecutive School Years--Cohort 6, Arithmetic Total Score Analysis

Variable		Consistent Underachievers		Consistent Overachievers	
Number of Students		64		71	
SES	$\bar{X}$	99.66		96.24	
	SD	8.87		8.24	
	g^1	0.02		1.00	
Sex[2] (girl=1;boy=0)	$\bar{X}$	0.52		0.61	
Ethnic Group[2] (white=1; non-white=0)	$\bar{X}$	0.79		0.94	
		Analysis Year		Analysis Year	
		SP72	SP73	SP72	SP73
Pretest (Nat'l Norm =476 for SP72; Interpolated Nat'l Norm=509.5 for SP73)	$\bar{X}$	492.63	469.79	491.72	553.62
	SD	46.18	42.71	45.24	47.82
	g	0.07	0.60	-0.70	-0.65
Posttest (Interpolated Nat'l Norm=509.5 for SP72; Nat'l Norm =545 for SP73)	$\bar{X}$	469.79	450.80	553.62	609.94
	SD	42.71	43.08	47.82	46.58
	g	0.60	0.72	-0.65	-0.68
Posttest Residualized for Pretest and SES	$\bar{X}$	-45.95	-43.04	40.16	39.74
	SD	30.77	26.32	15.03	17.45
	g	-2.40	-2.64	0.84	1.08
Attitude toward Math at Time of Pretest	$\bar{X}$	97.87	97.13	102.77	104.86
	SD	10.58	10.78	8.49	8.98
	g	0.24	0.24	-0.28	-0.57

Note: Blank entries indicate no data collected. (continued)

[1] Index of skewness

$$g = \frac{\frac{\Sigma x^3}{N}}{\frac{\Sigma x^2}{N}\sqrt{\frac{\Sigma x^2}{N}}}$$

, where g = skewness index,

x = deviation of a score from $\bar{X}$, where $\bar{X}$ is that group's mean on variable X, and

N = number of students.

[2] For binary coded variables the $\bar{X}$ is equal to the proportion of students coded with a "1."

TABLE V-53 (continued)

Variable		Consistent Underachievers		Consistent Overachievers	
		SP72	SP73	SP72	SP73
Attitude toward Math at Time of Posttest	$\bar{X}$	97.13	96.85	104.86	103.76
	SD	10.78	9.09	8.98	8.57
	g	0.24	0.37	-0.57	-0.44
Utilization of Objectives	$\bar{X}$	2.84	2.83	2.92	2.51
	SD	0.50	0.41	0.35	0.63
Individualization in Decision Making	$\bar{X}$	1.27	1.20	1.23	1.06
	SD	0.31	0.25	0.28	0.18
Teacher or Locally Developed Materials	$\bar{X}$	2.51	2.66	2.73	2.42
	SD	0.65	0.48	0.54	0.55
Individualization of Instructional Pace	$\bar{X}$	2.17	2.47	2.38	1.94
	SD	0.42	0.64	0.49	0.89
Scheduling Characteristics	$\bar{X}$	1.63	1.44	1.34	1.25
	SD	0.82	0.58	0.71	0.57
Use of Performance Agreements	$\bar{X}$	2.39	2.72	2.21	1.66
	SD	0.86	0.67	0.92	0.90
Classroom Group Organization	$\bar{X}$	1.94	2.30	1.82	1.83
	SD	0.68	0.65	0.72	0.63
Teaching Unit Composition	$\bar{X}$	2.33	1.59	2.50	1.46
	SD	0.69	0.61	0.62	0.63
Completeness of Instructional Package	$\bar{X}$	2.65	2.54	2.80	2.00
	SD	0.42	0.55	0.36	0.84
Utilization of Student Evaluation	$\bar{X}$	2.77	2.56	2.86	2.16
	SD	0.54	0.73	0.41	0.83
Level of Innovation	$\bar{X}$	22.48	22.30	22.81	18.30
	SD	2.22	3.13	1.44	4.60

(continued)

TABLE V-53 (continued)

Variable		Consistent Underachievers		Consistent Overachievers	
		SP72	SP73	SP72	SP73
Degree of Individualization	$\bar{X}$	8.60	8.95	8.68	6.82
Number of School Days per Year	$\bar{X}$	183.75	184.25	181.13	181.72
	SD	7.81	7.66	4.61	4.64
Number of Minutes per Day	$\bar{X}$	55.30	53.44	52.11	53.24
	SD	5.77	5.65	6.67	7.46
Teacher/Student Contact Ratio[3]	$\bar{X}$		3.47		3.61
	SD		0.49		0.56
Number of Teachers	$\bar{X}$	2.44	1.33	2.45	1.32
	SD	1.34	0.73	1.39	0.87
Teacher Age	$\bar{X}$	37.24	35.76	37.71	38.75
	SD	8.08	8.88	8.80	9.60
Teacher Socioeconomic Background	$\bar{X}$	96.50	96.56	95.70	94.11
	SD	5.84	8.34	5.78	9.15
Teaching Qualifications	$\bar{X}$	98.45	100.21	99.78	105.06
	SD	6.03	8.14	7.46	7.48

[3] Relative measure of teacher/student contact equal to the percentage of teacher time per individual student.

second analysis year, the overachievers had a much lower Level of Innovation and a noticeably lower Degree of Individualization than did the underachievers. The underachieving group had about the same Level of Innovation and Degree of Individualization in both years; the overachievers underwent a decrease on these two dimensions from 1972 to 1973.

The characteristics of the teachers of these two groups during the seventh grade were very similar (i.e., during the 1971-72 school year). In the eighth grade, however, the overachievers had older, more qualified teachers than did the underachievers.

Looking over the patterns of instruction received by the consistent overachievers across all three cohorts, there is no evidence to support the hypothesis that Level of Innovation is positively related to arithmetic achievement. For all three cohorts, the students who consistently overachieved in both years received arithmetic instruction characterized by a similar or noticeably lower Level of Innovation than that received by the consistent underachievers. This same general finding applies as well to our superscale, Degree of Individualization.

One additional finding to be noted is that the largest differences between the SES distributions of consistent overachievers and underachievers occurred in the Cohort 6 analysis for arithmetic/mathematics. Both the mean and the skew of the SES distributions were noticeably different. Again, although the generalizability of these findings is limited, the lower mean and larger positive skew of SES in the Cohort 6 consistent overachievers (relative to the underachievers) may mean that consistent overachievement occurred more frequently in lower SES students than in higher SES students in this analysis sample.

An examination was also made of the groups of students who had unusual but inconsistent patterns of mathematics achievement to see if their patterns of treatment changed systematically from year to year. However, no clear findings emerged.

VI. SUMMARY AND CONCLUSIONS

A. Study Objectives

Project LONGSTEP was designed to assess, across a three-year period, the overall impact of intensive, innovative educational practices on student outcomes. The implicit hypothesis leading to this research was that elementary and secondary schools implementing large-scale, intensive innovation (as defined by a select set of educational practices) would achieve substantially greater improvement in student performance than would more traditional schools.

Specifically, the objectives of the study were as follows:

1. to design a system for studying the educational environments of students in terms of meaningful dimensions which underlie various educational approaches or treatments;
2. to establish a large-scale data base consisting of treatment characteristics, student characteristics, and educational outcomes for a select sample of educational programs involving intensive and highly innovative educational practices;
3. to determine, as comprehensively as possible over a three-year period of time, the impact of intensive innovation upon student performance on standardized achievement tests and on measures of educationally relevant attitudes; and
4. to attempt to identify the dimensions of educational components present in a select sample of highly intensive, innovative programs that exhibit the greatest impact on student outcomes.

The initial design for Project LONGSTEP was developed in 1969, implemented on a limited basis, and modified during the 1969-70 school year. Full implementation began during the 1970-71 school year.

B. Sample Description

During the early stages of Project LONGSTEP, a nationwide search was undertaken for educational programs incorporating intensive, innovative practices. Guidelines for program selection emphasized program scope and intensity, instructional content, anticipated program continuation, and willingness to cooperate in a multiyear study. In general, these guidelines reflected the desire to include in the study innovative educational strategies (as of the 1969-70 era) representing sizable investment of school resources and having impact on a significant proportion of students at different grade levels in each school.

Through a combination of intensive literature search, interviews with program staff, outside consultant review, and site visiting, schools located in 14 school districts were selected and agreed to participate. Of the districts selected, 13 participated throughout the entire implementation phase and provided the data on which this report is based. These districts were located in nine states: California, Florida, Kentucky, Minnesota, Pennsylvania, Texas, Utah, Washington and Wisconsin. Altogether, some 30,000 students, 80 schools, and 1,500 teachers participated in the study during its three years of implementation.

The 13 districts included in Project LONGSTEP provided variability along various dimensions, as illustrated by a number of school and community characteristics documented during the 1970-71 school year. The communities served by the school districts ranged from 2,500 to over 600,000 in population and varied from rural to urban-metropolitan in setting. Their diversity in socioeconomic level is notable, as evidenced by the percentage of students in the participating schools receiving free or reduced-price lunches. This percentage ranged from less than 1 percent to nearly 30 percent. The reported instructional cost per pupil in the participating school districts varied from a low of $540 to a high of $1,050. Another indicator of the diversity in the sample was the percentage of minority members in the participating schools; this ranged from less than 1 percent to over 30 percent.

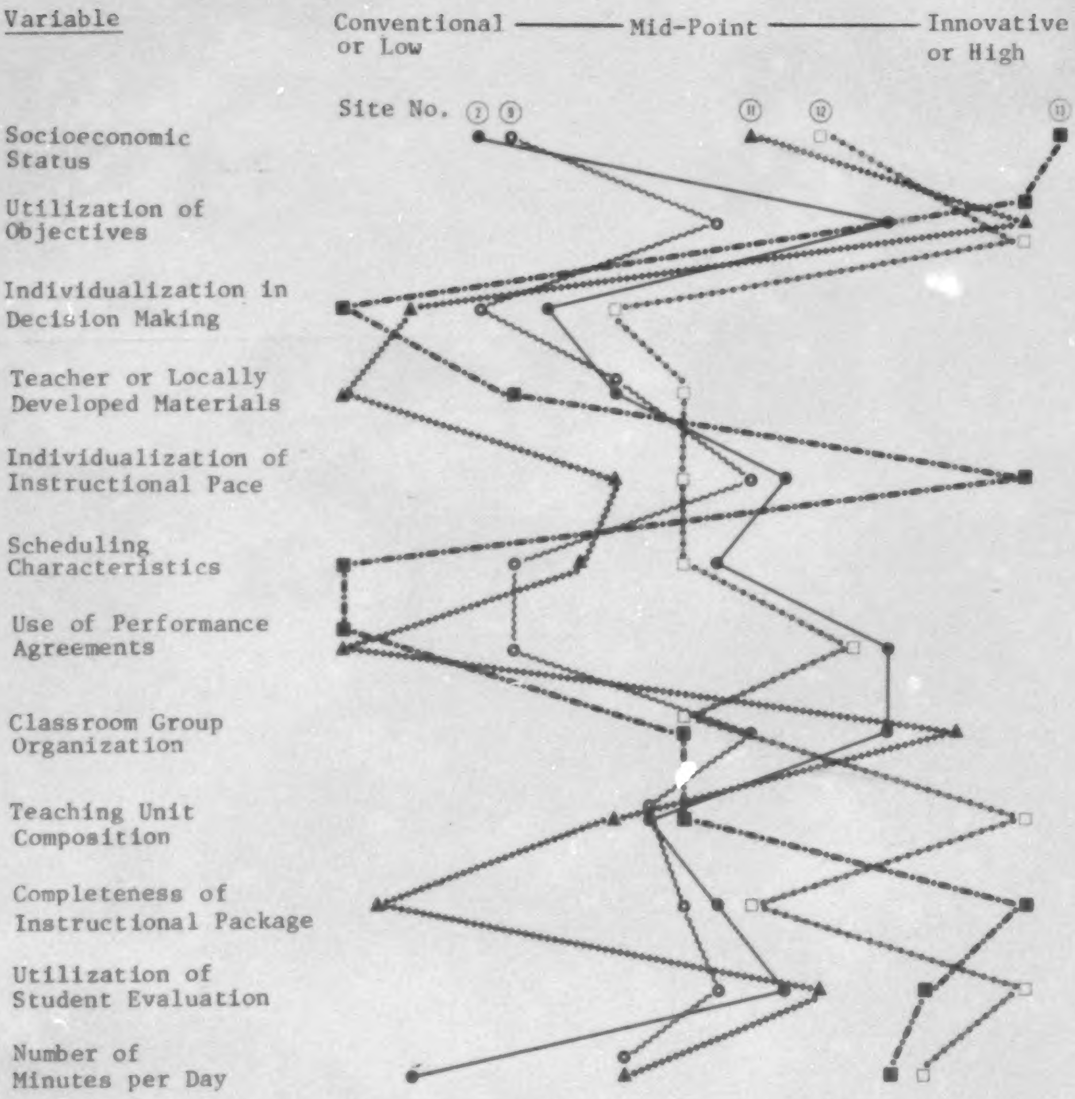
Although the study spanned grades 1-12, not all grades in each district were included. Nor did the study involve a random sample of instructional programs in American schools in general since such a sample would not have been likely to contain the diversity or the intensity of innovation necessary to achieve the study's objectives. Therefore, our selection criteria focused primarily on the identification of innovative schools. To ensure diversity, however, an effort was also made to include a range of educational practices, from the conventional to a highly diversified collection of innovations. This was done in two ways. First, all students in a selected grade in each participating school were included in Project LONGSTEP, regardless of whether or not they were involved in an innovative program. Second, an effort was made to include, wherever feasible, a number of conventional schools in order to ensure adequate diversity in treatment variables. Thus, variations in educational practices, by design, occurred both within and across the participating districts and schools.

The educational innovations encompassed by Project LONGSTEP were those of particular concern during the late 1960s and early 1970s--most are still of concern today. These included team teaching, multimedia emphasis, unique school design, use of paraprofessionals, variations in scheduling, and teacher-developed materials, as well as independent study, student selection of materials, and a number of other practices typically associated with individualized instruction.

C. Documentation of Educational Practices

In order to investigate the relationships between educational practices and educational outcomes in a very diverse group of schools and programs, a system was needed for describing and quantifying the educational experiences that each student had during a school year. It became evident early in the development of this system that, although there were a limited number of labels used by school districts to describe their educational activities, in practice there was a wide disparity among educational treatments sharing the same descriptive label. For example, "modular scheduling" is a widely used term for what are in fact various configurations of class schedules. At the same time, programs bearing different labels often turned out to be very much alike. Thus, one district's program of "individualized instruction" could be much the same as another district's "nongraded" program. Since this study proposed to investigate the relationships between educational practices and various educational outcomes in a highly diverse group of programs and schools, it was deemed necessary to develop a system by which complex educational experiences could be described and quantified with respect to specific observable characteristics rather than on the basis of variously defined local labels. In other words, what was needed was a methodology for describing educational treatments on the basis of their common denominators. To meet these needs the Educational Experience Analysis Guide (EdExAG) was developed as a data-gathering instrument. The use of the EdExAG made it possible to locate the educational experiences of students on a continuum ranging from traditional to innovative, or in some cases simply to record the presence or absence of certain practices. The major purpose of the EdExAG was to ensure that all students in a school who were receiving essentially the same basic educational experiences would belong to the same treatment group and would be distinguished from students receiving different educational experiences, even though both groups of students may have been participating in the same school "program."

During the 1970-71 school year, this treatment documentation methodology led to the identification of 141 EdExAG groups; during the 1971-72 and 1972-73 school years respectively, 167 and 228 groups were isolated. These groups represented all the different kinds of educational treatments provided by the schools for students participating in Project LONGSTEP. One illustration of the diversity of our sample is shown by the characteristics of the five sites (school districts) containing language arts EdExAG groups for the third grade during the 1972-73 school year. Figure VI-1 shows the average scores for these five sites on ten educational treatment indices derived from items on the EdExAG, plus student socioeconomic status (SES) and the Number of Minutes per Day spent on language arts activities.



Note: Data were aggregated across treatment groups and schools within site to yield the mean values plotted here. In terms of the sample identifications used throughout the report, these data represent the Cohort 1, SP73 Reading Total Analysis Sample.

Figure VI-1. Profiles for language arts treatments for the five sites containing third grade students during the 1972-73 school year.

D. Scope of This Report

The analyses discussed in this report focus on the extent to which intensive innovation was associated with substantial differences in student outcomes. Because of time and cost constraints, the analyses conducted for this report involved only three cohorts of students: those who were first graders in 1970-71 (Cohort 1), those who were fourth graders in 1970-71 (Cohort 4), and those who were sixth graders in 1970-71 (Cohort 6). These cohorts were selected for initial analysis primarily because they would permit the examination of educational changes occurring during the primary years (for Cohort 1), the elementary years (for Cohort 4), and the transition years from elementary to middle or junior high school (for Cohort 6). A future report will provide the findings for the remaining cohorts.

The analytic scope of this report was also limited to the relationship between language arts or arithmetic/mathematics treatments and three cognitive dependent variables: the Reading Total, Language Total, and Arithmetic Total Scores from the Comprehensive Tests of Basic Skills (CTBS). Outcome measures such as student attitudes and other possible correlates of change in attitudes and/or cognitive achievement were not directly analyzed as dependent variables. Rather, these variables were considered during the analyses in which attributes of unusually effective or ineffective educational treatment groups and consistently high-achieving or low-achieving students were examined. An analysis of the generalized impact of innovation throughout the entire school environment (including social studies and science treatment data) will be presented in a future report.

The present report not only covers the substantive findings of Project LONGSTEP but also presents full details on the methodology used. Those readers particularly interested in methodology will perhaps want to devote considerable attention to (1) Chapter IV, which reviews the analytic procedures applied, and (2) Appendix IV-E, which describes the relationship between regression analysis involving multilevel predictor variables (e.g., student scores and indices of treatment-group attributes) and commonality analysis. The technical properties of regression analysis and commonality analysis are too complex for discussion in this brief summary chapter. Since they underlie a number of the procedures used for this report, however, the technically inclined reader is directed to Appendix IV-E for a discussion of these properties.

Project LONGSTEP represents a pioneering effort in educational evaluation. It is perhaps the first large-scale, longitudinal study to associate the educational attributes underlying school programs and the characteristics of teachers with only those individual students exposed to their influence. The vast amount of data collected is also notable--other studies may have collected information on more students, but for sheer quantity and complexity of data, this study is unique. When all the tests and questionnaires administered to students and teachers and all the Educational Experience Analysis Guide data collected on language arts, mathematics, social studies and science treatments over a three-year period are considered, some 3,500 individual items of data could be associated with a single student. When the scales or indices derived from each instrument are also considered, the total number of variables potentially available on each participating student (assuming no missing data) increases to approximately 4,000 variables per student. Because of the potential of such a large and detailed data base, a series of generalized data files and codebooks are being produced for the U.S. Office of Education so that these data can be made available to other researchers.

E. Findings in Review

Project LONGSTEP was designed to determine if intensive innovation is related to substantial gains in achievement by studying the correlates of longitudinal achievement growth in a diverse sample of schools, selected primarily because of their emphasis on intensive educational innovation. To accomplish this objective, the study assumed that if innovation does have a dramatic impact on achievement, then:

- variation with respect to intensity of innovative emphasis, even within a sample of generally innovative schools, should be positively correlated with achievement;

- such impact should be a general trend across different classrooms, schools and sites and should be consistently present in consecutive school years;
- the achievement gains demonstrated by Project LONGSTEP's generally innovative sample of schools should noticeably exceed the gains shown by the CTBS norm sample; and
- such impact should be large enough to be demonstrated by a number of different methods, including those involving the kinds of statistical adjustments normally required in analyzing nonexperimental research data.

Because of the associational, naturalistic research approach used by this project, it was deemed essential that the study's key hypothesis be evaluated from a number of slightly different methodological perspectives so as to minimize the possibility that findings would be highly method- or assumption-dependent. Conclusions could then be based on converging lines of evidence. For this reason, the analyses implemented for this report were, by design, as intensive as possible within project time and cost constraints. The methods used included:

- examination of pretest-to-posttest gains to ascertain the magnitude of educational growth across school years;
- comparison of mean pretest and mean posttest with CTBS national norms to assess the extent to which the educational growth shown by the average student in each Project LONGSTEP sample differed from published CTBS national norms;
- examination of the confounding among initial achievement level (measured by the pretest), student socioeconomic status (SES), treatment-group membership, and treatment attributes to evaluate the extent to which different students tended to be exposed to different treatment approaches;
- assessment of the extent to which initial achievement level (i.e., pretest), student SES, and treatment-group membership were overlapping predictors of posttest achievement by means of a technique called commonality analysis;

- assessment of the predictive overlap in a linear growth model of achievement, including the following as predictors of posttest performance: pretest, student SES, Level of Innovation (a superindex providing an overall measure of treatment innovative intensity with respect to the ten key process attributes shown in Figure VI-1), Number of Minutes per Day (a measure of the amount of class time spent on language arts or mathematics activities), and Teaching Qualifications (a measure of the experience of the teacher or teachers of each student);
- computation and examination of the regression coefficients for the predictors in the growth model (as well as the intercorrelations among posttest and the predictors);
- evaluation of the extent to which this predictive growth model was reasonably applicable to students at all SES levels or all pretest levels (the "moderator" analyses);
- inspection of the correlations (across analysis years and cohorts of students) between treatment attributes and a residualized achievement score equal to that part of each student's posttest score that could not be linearly predicted from his/her pretest and SES;
- identification and examination of treatment groups whose students, on the average, did much better or much worse than was expected on the basis of their pretest and SES;
- identification and examination of students who, in relation to their pretest and SES scores, showed substantial gains or conspicuous lack of gains in two consecutive school years;
- examination of another educational growth model in which the primary treatment variable was not Level of Innovation but Degree of Individualization, a superindex providing a measure of the extent to which the educational treatments involved a program-level emphasis on individualization of instruction; and

- evaluation of the extent to which trends were consistent across cohorts of students, pretest/posttest analysis years (1971/1972 growth, or 1972/1973 growth), analytic methods, and cognitive dependent variables.

The findings based on these analyses are discussed throughout the report. In the remainder of this section, we will summarize the findings by reviewing them in terms of four basic questions.

Were Students of Different Socioeconomic Status and Pretest Levels Exposed to Different Kinds of Treatments?

Although there were numerous statistically significant correlations between socioeconomic status (SES) or pretest level and the treatment attributes examined, there were no general, systematic trends across cohorts. This result was probably due, at least in part, to (1) the slightly different schools and sites represented in the three analysis cohorts and (2) differences in the treatment approaches to which the same cohort of students was exposed in the two school years. Although the most general finding is that treatment attribute correlations with student SES and pretest were specific to a particular cohort, to a single analysis year, or to a particular variable, a number of trends were noted. Even the most notable of these trends, however, was very small relative to the magnitude of the SES differences among treatment groups.

Lower SES students in the reading, language and arithmetic analyses tended (1) to be exposed to slightly less innovative treatments (relative to their respective samples) in terms of Level of Innovation or Degree of Individualization and (2) to be members of treatments that, relative to each LONGSTEP sample, allocated more time per day to language arts or mathematics activities. This pattern of results was most discernible in the Cohort 1 and 6 analyses for reading and language.

With the exception of the amount of language arts class time per day, it was difficult to discern any systematic associations between pretest and the treatment variables across cohorts. In general, lower pretest students in each sample tended to be exposed to more language arts instruction than the higher pretest students. Furthermore, treatment groups varied less with respect to initial achievement level (pretest) than they did with respect to mean SES level.

In terms of the confounding of treatment-group membership and the treatment attributes, the findings noted above can be summarized in the following ways:

- patterns of confounding differed across analysis samples, and
- students with lower SES and pretest tended to be members of schools and instructional programs that spent more time on language arts and math activities than did programs with higher SES and pretest-level students.

Because the relationships between SES or pretest and the treatment variables varied across analysis samples and cohorts, it can also be noted that if statistically adjusting out pretest and SES differences did result in any regression artifacts in the analyses discussed in this report, it is not likely that these artifacts would be similar in all analyses.

To What Extent Is Achievement Related to Pretest and Socioeconomic Status?

Both initial achievement level or pretest and student socioeconomic status (SES) are usually substantially related to posttest achievement in school effects studies. The results of Project LONGSTEP are no exception. It should be noted, however, that pretest was a vastly better predictor of posttest than was student SES. Furthermore, the predictive value of SES was reduced almost to zero in the Project LONGSTEP analysis after SES overlap with pretest was taken into account.

The relationship between pretest and posttest, however, was still very large even after student differences with respect to SES and after group differences with respect to mean pretest level were taken into account.

Although it was the treatment correlates of posttest achievement that were of primary interest in Project LONGSTEP, these findings are of value to educators and researchers. Based on a diverse sample of schools, these findings suggest that when pretest is known, it is not necessary to use the SES level of a student or of a treatment group (or school) as a predictor of individual student posttest achievement. These findings also indirectly imply that SES cannot be expected to serve as a reasonably accurate student-level or school-level substitute predictor of student pretreatment achievement level in survey research studies.

Did Intensive Educational Innovation Result in Substantial Growth in Achievement?

As noted throughout this report, the most acceptable approach to analyzing data obtained by means of nonexperimental/associational research, like Project LONGSTEP, is to use a number of different analysis methods and look for consistent results across methods. Therefore, Project LONGSTEP analyzed:

(1) pretest and posttest gains;

(2) treatment-group differences in achievement growth;

(3) a growth model in which growth in achievement was related to treatment attributes;

(4) a posttest score residualized on the basis of pretest and SES;

(5) treatment groups whose students, on the average, performed differently than was expected from their pretest and SES; and

(6) consistently overachieving and underachieving students.

Findings with respect to the impact of educational innovation obtained by means of these methods can be organized into two general categories. The first group of findings discuss the overall amount of educational growth demonstrated by each of the analysis samples and compare this growth with that of the CTBS norm sample. The second group of findings show how variation in innovation among the treatment groups present within each analysis sample was related to variation in achievement growth. The second group of findings are summarized separately for each of the three cognitive dependent variables--reading, language and arithmetic.

Growth in achievement and norm comparisons. Comparisons among mean pretest, mean posttest, and CTBS national norms showed that:

- the greatest gains in reading, language and arithmetic achievement occurred in the earlier grades;
- the magnitude of yearly gains in achievement was inversely related to grade level;
- for all grades, achievement gains seemed to be larger in arithmetic than in language arts (reading and language);
- although the gains were not dramatic, achievement growth during the third grade was somewhat greater than that in the CTBS norm sample; and
- the Project LONGSTEP posttest means were not conspicuously farther from their published norms than the pretest means were from their norms.

Furthermore, it is possible that the gains shown by these LONGSTEP samples may have overestimated growth rather than underestimated growth because only those students with no missing data on key analysis variables (i.e., those indices included in the growth models examined) were included in the samples analyzed. It seems reasonable that growth rates might be higher for these students than for students who were not included in the analysis samples because students with no missing data had higher pretest scores (on the average) than students with missing data and because students with higher pretest scores probably have higher growth rates.

In any event, the hypothesis that students enrolled in fairly intensive, innovative educational programs would (on the average) exhibit substantial yearly gains in achievement is not supported by these findings.

Reading achievement. The findings for the analysis of reading achievement were as follows:

- The proportion of criterion variance, accounted for solely by treatment-group differences, independent of the effects of student socioeconomic status (SES), initial achievement level (as measured by pretest), and Teacher Age, SES Background and Teaching Qualifications, is large for second and third graders, less for fifth and sixth graders, and almost nonexistent for the seventh and eighth graders. When considered along with overall pretest/posttest gains and the norm comparisons discussed previously, these findings show that (1) meaningful but not dramatic differences in educational growth did exist among Project LONGSTEP treatment groups, especially those involving students in the early elementary grades, and (2) treatment impact was inversely related to grade level.

- Both the irregularity of the overall pattern of regression coefficients and the small uniqueness for Level of Innovation plus Number of Minutes per Day in all samples but one suggest that no consistent overall relation existed between innovative intensity (as indexed by Level of Innovation) and posttest reading performance, or between individualization emphasis (as indexed by Degree of Individualization) and posttest reading performance. In fact, in the only series of analyses in which the treatment variables were related to a fairly noteworthy amount of educational growth (for the students in Cohort 1, during grade 3), the regression coefficients for Level of Innovation and Number of Minutes per Day were negative. The scale Teaching Qualifications was not related to posttest performance to any meaningful degree. The part-correlations in which posttest was adjusted on the basis of pretest and SES also did not suggest that consistently positive and substantial associations of innovative intensity with posttest reading achievement were present across analysis samples. Since the same general lack of meaningful impact was observed whether or not the criterion or the predictors were residualized, the hypothesis that a higher Level of Innovation is positively and consistently related to student reading achievement is not supported.

- The analysis of unusually effective treatment groups in reading provided no clear evidence that Level of Innovation or Degree of Individualization was consistently related to posttest performance. For example, for two of the cohorts, the high-performing groups in 1972 were characterized by greater innovation and greater individualization than the poorer performing groups. In 1973, however, the high-performing groups had lower innovation and less individualization. For the other cohort, the positive outliers in 1972 differed little from the negative outliers on these dimensions, but the positive outliers in 1973 demonstrated a greater Level of Innovation and a higher Degree of Individualization.
- In general, the consistent overachievers in reading tended to be members of programs with a lower Level of Innovation or Degree of Individualization than the consistent underachievers. The absence of strong and consistent positive relationships clearly does not support the hypothesis that Level of Innovation or Degree of Individualization (as defined here) is positively related to reading achievement. Perhaps the most notable finding in this analysis was the fact that Cohort 1 consistent overachievers on the CTBS reading test, on the average, spent considerably more time per day on language arts during the second grade than did underachievers--113 minutes compared to 85 minutes. As third graders, however, the overachievers were exposed to about 20 minutes of instruction per day less than during the previous school year and yet, again demonstrated larger than expected gains in reading achievement. These results suggest that increased exposure to language arts during the second grade for some students may significantly improve their chances of demonstrating substantial gains in reading achievement, even in later grades. Another interesting finding was

that some of the largest differences between overachieving and underachieving students, across all grades, occurred with respect to the Reading Interest scale. The overachievers consistently displayed a higher level of Reading Interest during both the pretest and posttest school years.

Language achievement. Although reading and language skills are conceptually distinct and may be developed by means of different instructional materials and activities (e.g., reading exercises versus grammar exercises), the skills assessed by the CTBS Reading Total and Language Total Scores are such that these two measures tend to be fairly highly correlated. Thus, the findings for language achievement that are summarized next were quite similar to those for reading achievement:

- As in the reading achievement analyses, with the exception of achievement growth during the third grade (i.e., Cohort 1, SP73 analysis), only a small proportion of the predictable variation in posttest language performance could be uniquely attributed to the superscales, Level of Innovation and Degree of Individualization. The language achievement results with respect to Level of Innovation and Degree of Individualization demonstrated no trend for the regression coefficients for either of these superindices to be statistically significant and of the same sign across all six analysis samples, or across each pair of analysis data sets for a given cohort. The small proportion of variance in the language posttest performance accounted for by the treatment variables and the lack of consistent positive findings, then, clearly do not support the hypothesis that Level of Innovation or Degree of Individualization (as defined here) is related to increased language achievement.
- As was true in similar analyses for reading, the results of the analysis of unusually effective treatment groups were obviously conflicting. For no grade were the results consistent over the two years, nor were the results

consistent in any single year across all grades. In comparison with the poorer performing groups, the high-performing treatment groups were just as often characterized by a lower Level of Innovation and a lower Degree of Individualization as they were by a higher score on these superscales. The results for Teaching Qualifications, however, were not quite so contradictory. At least in the analyses where the positive outliers differed from the negative outliers on this scale, the teachers of the high-performing EdExAG groups tended to possess higher Teaching Qualifications. In no instance were the teachers of the positive outlier groups obviously lower in Teaching Qualifications.

- The average Level of Innovation and Degree of Individualization among the consistently overachieving students tended to be slightly lower than among the consistently underachieving students. In no case was the Level of Innovation notably higher among the overachievers. Also, the Cohort 1 overachievers in language (as they did in reading) spent, on the average, more time per day in language arts activities during the second grade than did the underachievers.

Arithmetic achievement. Findings for arithmetic achievement may be summarized as follows:

- Meaningful posttest differences existed among treatment groups, independent of concomitant differences with respect to SES, pretest and the teacher variables, especially in the lower grades. These findings, as well as those discussed earlier with respect to pretest/posttest gains and CTBS national norm comparisons, suggest that treatment impact (as measured by the CTBS Arithmetic Total Expanded Scale Score) was inversely related to grade level.

- Although treatment groups differed with respect to growth in achievement, the majority of all analyses showed that both Level of Innovation and Degree of Individualization (as defined here) were negatively, rather than positively, related to growth. This appeared to be the most notable trend in the samples taken as a whole. Therefore, the hypothesis that a higher Level of Innovation or a greater Degree of Individualization is positively and consistently related to student achievement in arithmetic is clearly *not* supported by our data. The results for Number of Minutes per Day and Teaching Qualifications did, however, show some tendency to be positively related to student achievement in arithmetic in some analysis samples.

- The analysis of unusually effective treatment groups also showed that effective and ineffective groups did not consistently differ with respect to Level of Innovation and Degree of Individualization.

- The patterns of instruction received by the consistent overachievers across all three cohorts did not yield any evidence to support the hypothesis that Level of Innovation was *positively* related to arithmetic achievement. For all three cohorts the students who consistently overachieved in both years received arithmetic instruction that was very similar or somewhat lower relative to Level of Innovation than that received by the consistent underachievers. The same general finding applied as well to the superscale, Degree of Individualization.

Trends across language arts and arithmetic. Summarized below are the major findings which appear to cut across reading, language and arithmetic.

- Pretest-to-posttest gains in the Project LONGSTEP samples were not notably better than those shown by national norms.

- No evidence could be found that either of the major treatment variables--Level of Innovation and Degree of Individualization--was substantially and positively correlated with posttest performance (regardless of whether or not posttest performance or the treatment variables were "corrected" statistically for differences in pretest and SES level).

- An analysis of the treatment attributes of unusually effective or ineffective treatment groups also could not identify treatment characteristics that consistently differentiated among such groups.

- Preliminary analyses showed that, on the average, findings regarding Level of Innovation and Degree of Individualization applied equally well to students at different SES or pretest levels.

- Meaningful differences in the teacher qualifications of groups performing at a higher or lower level than predicted were infrequent. Whenever they did occur, however, there was a slight tendency for higher teaching qualifications (as indicated by such items as education, teaching experience, certification and type of appointment) to be associated with better student performance.

- Treatment differences between students whose achievement, for two consecutive school years, greatly exceeded that expected on the basis of their pretest level and socioeconomic status showed that overachievers tended, if anything, to be members of programs characterized by a more moderate Level of Innovation or Degree of Individualization; while underachievers, on the average, tended to be in more innovative programs. Underachievement or overachievement was not noticeably more prevalent in high or in low SES students.

- The overachievers in Cohort 1 showed a rather discernible trend in the language arts analyses. As a group, they were exposed to more instruction per day during the second grade than were the underachievers. As third-graders, the overachievers (on the average) spent less class time per day on language arts than they did the previous year, and yet continued their pattern of overachievement. The underachievers spent the same, or slightly more, class time on language arts than did the overachievers in grade 3 but continued to fall behind. Although conclusions based on these highly select students can only be speculations, these results do suggest that early emphasis (that is, during the second grade) on language arts may facilitate exceptional growth in reading and language skills, at least in some students.

- Overachievers in arithmetic were *not* exposed to more arithmetic instruction (on the average) than were the underachievers in any of the three cohorts. The findings suggest, therefore, that the speculation regarding possible impact of increased instructional time cannot be extended to exceptional achievement in arithmetic.

- The Number of Minutes per Day spent on language arts was noticeably higher for children in the earlier grades (grades 2, 3 and 5) than for students in grades 6, 7 and 8. As an illustration of this difference, the students in grade 2 spent approximately 97 minutes per day on language arts as compared with 51 minutes per day spent by eighth-graders. In contrast, the Number of Minutes per day spent on arithmetic was relatively constant across all grades and ranged from approximately 51 to 53 minutes per day. Nevertheless, growth in achievement in the early grades exceeded similar growth in older students in reading, language *and* arithmetic. Project analyses did not suggest any explanations for these findings.

Can Schools Have a Substantial Impact on Educational Achievement?

The findings show that substantial educational growth was not uniquely associated with innovative emphasis at the treatment-group level. Even small amounts of growth were not positively and consistently related to our superindices, Level of Innovation and Degree of Individualization. It should not be concluded, however, that substantial growth was absent in the schools and students participating in Project LONGSTEP. On the contrary, as discussed previously, a number of educational treatment groups were identified whose students, on the average, did much better than was predicted from their pretest and SES level. For example, as indicated in Figure VI-2, the average achievement gains of sixth grade students in the most effective reading treatment groups differed dramatically from those exhibited by students in the groups that did much worse than was predicted from their pretest score and their SES level. The students in the poorer performing reading groups shown in Figure VI-2 had average pretest scores ranging between the 56th and 74th percentiles; their average posttest scores dropped slightly, ranging between the 41st and 66th percentiles. In contrast, the sixth grade students in the high performing reading groups, who had average pretest scores ranging between the 48th and 69th percentiles, improved noticeably, with average posttest scores ranging between the 63rd and 79th percentiles. Other examples of dramatic gains in achievement were shown in the analyses identifying individual students who, for two consecutive years, demonstrated posttest achievement differing greatly from that expected on the basis of their pretest and SES scores.

The outlier treatment-group analyses and the consistent underachiever/overachiever analyses do provide evidence that large gains in cognitive achievement were occurring in Project LONGSTEP. Thus, even though the overall project findings showed that dramatic school effects were not associated with intensity of innovative emphasis, the findings discussed here suggest that substantial achievement gains, over and above those expected on the basis of pretest and SES, do occur in public schools. It should also be remembered that the overall analyses showed that small but

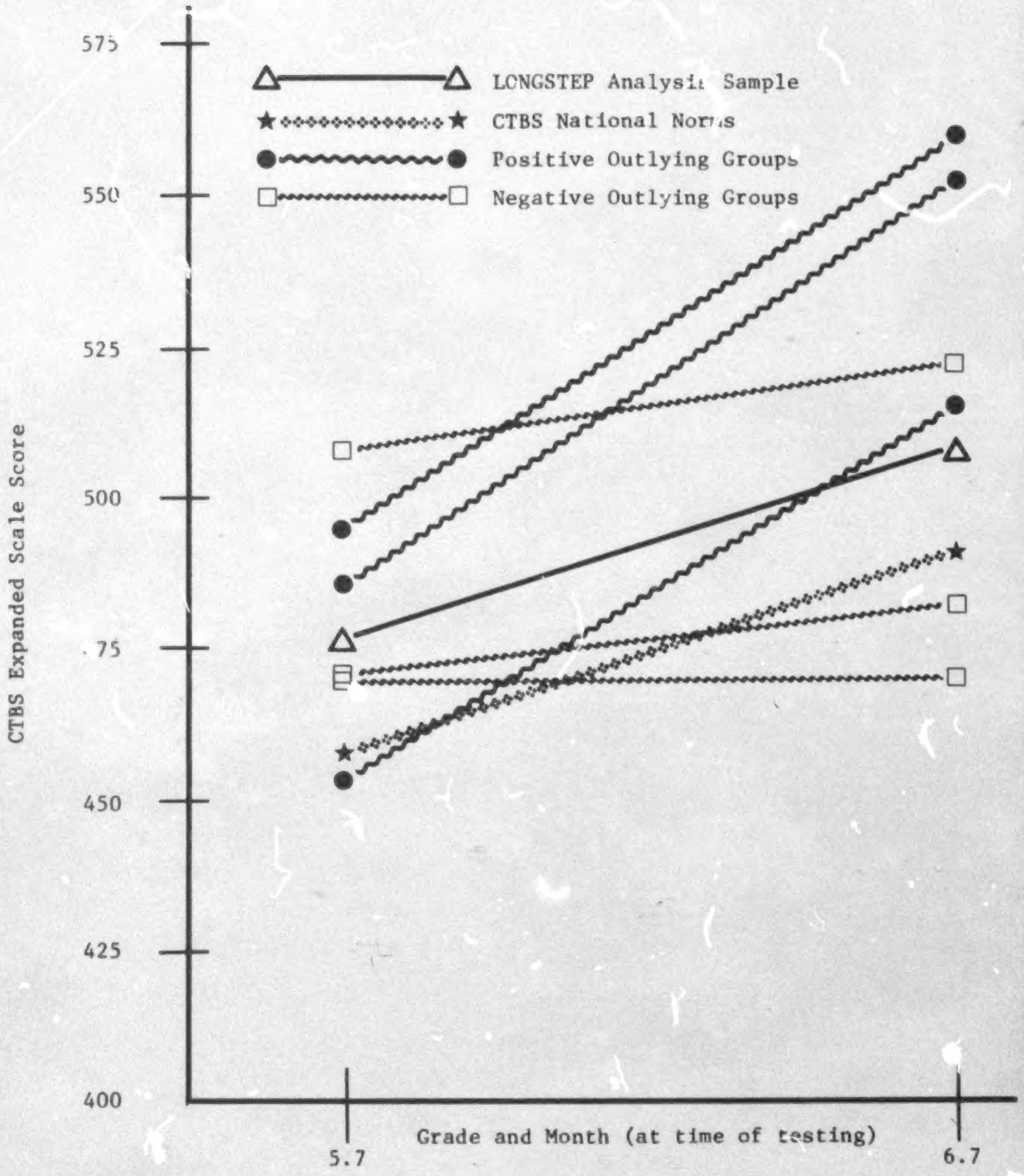


Figure VI-2. Cohort 4 treatment groups where students achieved more or less than was expected from their pretest and SES level (Spring 1973 Reading Analysis).

meaningful posttest differences existed among treatment groups, although they were not uniquely associated with the project's measures of innovative intensity.

These findings certainly are not surprising to teachers, principals, parents or even researchers who have personally witnessed impressive growth in cognitive skills. They are, however, somewhat more "newsworthy" in terms of educational evaluation results. Taken together, then, these findings must lead us to conclude that:

- different educational treatments and schools did produce meaningful differences in achievement (especially in the early elementary grades); and
- although substantial gains in achievement were not demonstrated by a majority of Project LONGSTEP treatments/schools and students, dramatic improvements in reading, language and arithmetic skills were found in some schools.

Future research and evaluation efforts studying the correlates of achievement should be directed at identifying the antecedents of such significant educational gains.

F. Study Limitations

Because of Project LONGSTEP's associational/naturalistic design and new approach (as of 1969-70), a number of factors must be considered when interpreting the results obtained. Although many of the technical problems that can beset a research venture of this magnitude and nature have been addressed by the study, such factors as its necessary associational design, state-of-the-art limitations in psychometrics, and practical time and cost constraints have resulted in limitations of which the reader should be aware. This section will briefly review a few of the limitations that pertain to Project LONGSTEP that should be considered in the design and implementation of future large-scale studies of this kind.

It should be noted, however, that probably none of these limitations severely compromises the study's basic findings--especially since the major interest of the study was the presence or absence of substantial improvements in achievement. Nevertheless, such limitations do make interpretation of the results more difficult and leave some findings open to more than one possible interpretation. For this reason, straightforward interpretations are not always appropriate, and conclusions must be based both on converging lines of evidence and on judgments as to the validity and reasonableness of the findings. The issues noted in the following paragraphs are not unique to this study, but they are ones we believe are important to consider in interpreting our findings.

1. Perhaps Project LONGSTEP's greatest weakness is one that it shares with almost all other large-scale educational evaluation projects conducted in public schools--that is, the necessity of using an associational/survey research design. The inability to conduct controlled studies employing random assignment of students or classes to defined treatment conditions is well-known and is not likely to change in future years. Schools, after all, are community service organizations, not laboratories. The understandable desire of each community and its local school personnel to provide the kind of educational treatment they believe is best for their students certainly doesn't allow much latitude for experimentation. Quasi-experimental designs are available when and where limited control is feasible. Even if some kind of initial control were possible, however, gradual

changes in the nature of the treatments across time as well as natural changes in student groups due to attrition would tend to destroy the controlled nature of any research design for a longitudinal study. (Such an approach would, of course, be valuable if the primary process being studied were change in the treatment itself.)

The major disadvantage of nonexperimental research designs is that they cannot yield straightforward conclusions of the sort that can be obtained by means of "orthogonal" experimental designs. In a true experiment (a) the hypothesized antecedents of student achievement can be specified in such a way that their separate or joint (i.e., interaction) effects can be independently estimated, and (b) alternative explanations for observed effects can be minimized by equalizing the impact of such "nuisance" factors across groups through random assignment of students to the various treatment conditions. Without such control (as in associational research efforts), the possible antecedents of achievement are likely to be correlated or confounded with one another in known and unknown ways.

When known factors are confounded, it is not possible to estimate their separate impact unambiguously, simply because such factors do not occur separately in the sample analyzed. For example, in Project LONGSTEP, the conceptually distinct attributes of educational treatments indexed by the Educational Experience Analysis Guide tended to occur together in the same schools and were therefore highly confounded. Thus, estimating their separate effects was not feasible in these analyses. Similarly, the confounding of student SES and pretest with Level of Innovation meant that their separate effects on posttest could not be completely disentangled. The presence of such effects, of course, are findings in their own right since they describe processes that are present in ongoing educational institutions. However, confounding of treatment with student attributes can be particularly troublesome since one very likely explanation for outcome differences is that such differences existed prior to the treatment. To be technically correct, statistically adjusting out preexisting differences is logically inadequate because the reason that groups differ prior to treatment is unknown--that is, differences among groups prior to the treatment may be due to the fact that group membership is confounded with some unknown variable or variables. These unknown variables, then, and not the treatment, may be the real reason for the "adjusted" posttest differences among the groups. Because such unknown antecedents are always present in associational designs, definitive findings are not obtained. However, since survey research will undoubtedly continue to be the major source of information upon which national educational policy is based, rigorous and intensive analyses of such data must be combined with professional judgment so that educationally valid conclusions may be reached. In short, conclusions from nonexperimental research must be based on (a) trends shown by a number of different analytic methodologies involving different assumptions, (b) judgments regarding the educational processes being studied, and (c) knowledge of the characteristics of the students exposed to the processes studied.

It should be pointed out that not all the possible consequences of nonexperimental research are disadvantageous. One very significant advantage is that, by using naturally-occurring groups, it is possible to avoid the artificiality created by pure experiments. An experimental study, for example, might give very clear-cut results from which a limited number of conclusions can be drawn. However, these conclusions might not be applicable when the educational treatment is applied operationally in a naturalistic school setting.

2. Another limitation that is present in virtually all longitudinal studies concerns the choice of measurement units. Analyses involving (a) comparisons of simple pretest/posttest gains, (b) residualized gains (posttest residualized on the basis of pretest and SES), or (c) linear growth models including pretest as a predictor, all require that the test scores involved represent interval scales--or, at least, scores which "are only mildly distorted versions of an interval scale" (Gardner, 1975, p. 55). The reason for this requirement, of course, is that changes or differences across time are the major outcomes of interest, and differences at one point in the scale should be fairly comparable to those at another point, a property that is present in interval but not ordinal scales. (This discussion should not be confused with the more substantive issue that growth rates may differ for students at different pretest achievement levels.)

The interval scale requirement is undoubtedly fairly crucial in nonexperimental research designs like that of Project LONGSTEP where different treatment groups often have different pretest levels and where growth demonstrated by such disparate groups is then related to treatment attributes.

The ideal solution to this problem is for all longitudinal studies to utilize tests for which interval score scales are available. Unfortunately, none of the tests available on the market has score scales that are demonstrably and unambiguously true interval scales. Project LONGSTEP chose to use the CTBS Expanded Standard Scores provided by the test publisher since these scores were designed to constitute an interval scale. Although not perfect in this respect, they are probably close enough to a true interval scale that any conclusions based on procedures requiring their use are reasonably sound.

3. A third complicating factor was the tendency for the nature of the treatments present in participating schools (as indexed by the study's Educational Experience Analysis Guide scales) to change across time. There was even some evidence to suggest the presence of a longitudinal trend toward less innovation (see Appendix IV-A). Although it is possible that these scales were longitudinally unstable indices of treatment attributes, this is not likely since (a) the same project personnel usually visited the same school during the course of the study, and (b) treatment documentation for the 1972-73 school year was accomplished by noting changes from the previous year. It is more likely that changes were under- and not overestimated by the 1972-73 treatment scales.

Intensity of innovation was not found to be related to substantial gains in achievement nor to small consistent gains across grades and years. It seems possible that one of the primary reasons why consistent effects were not observed, especially during two consecutive school years for the same cohort of students, may have been due to longitudinal fluctuations in treatment approaches or to student movement in and out of fairly stable treatments. In either case, exposure to different treatment approaches in consecutive school years may not have been beneficial to students. If this hypothesis is true, the effects of a mixture of weak—beneficial and weak-ineffective (or harmful) approaches in consecutive school years would tend not only to lead to small effects for one school year but would also tend to minimize longitudinal gains in achievement.

On the other hand, Project LONGSTEP's inconsistent results across school years may have been due to the lack of dramatic effects in the data and slight regression artifacts produced by statistically adjusting for different patterns of SES, pretest and treatment confounding in the data for consecutive school years or for different cohorts of students.

Limitations such as these are not likely to hide the dramatic achievement gains which were of concern to Project LONGSTEP. However, they may attenuate the longitudinal effects of schooling and limit the ability of survey studies, like the present one, to identify small but consistently beneficial treatment components.

4. A fourth limitation of Project LONGSTEP and almost all large-scale studies like it was the necessity imposed, oddly enough, by one of its chief strengths--its magnitude. Because of the large number of students (approximately 30,000), teachers (approximately 1,500), schools (approxmately 80), and grades (1 through 12), as well as the large amount of data gathered in a three-year period, it was necessary to analyze the data primarily in terms of overall statistical tendencies, rather than focusing intensively on any one program or group. Thus, there could conceivably have been important results for a particular innovative program that were not identified because they were buried by a mass of data not showing those results on an overall basis.

In order for the possible presence of such effects to be examined efficiently and related to the various treatment measures analyzed for this report, the attributes of outlying treatment groups were inspected. As noted previously, these analyses showed that certain of the treatment groups participating in the study did show dramatic achievement gains and losses but also that such effects were not clearly and consistently related to the same treatment attributes.

5. Another limitation, which Project LONGSTEP probably shares with school effects studies in general, was that it was easy to lose sight of

the individual student because of the necessity of looking for overall, average trends. A mean is the best single descriptor of a group--in regard to a particular characteristic--but it is not necessarily a good descriptor of every individual in that group, nor is it equally good in all groups (because of different within-group distributions).

Some innovative methods may work extremely well for a small percentage of students, work counterproductively for another small percentage of students, but make little difference to the rest--the majority of students. In this case, the effect on the students for which the procedure was particularly effective would be offset by the effect on the students for which the procedure was counterproductive. And even if there were no students for whom the procedure was counterproductive, the results for the very small group for whom it was extremely effective would be masked by the results for the vast majority for whom it made no difference.

Although Project LONGSTEP was not designed to identify effective trait/treatment combinations, the analysis of consistently overachieving and underachieving students did attempt to see if students who did very well in school were exposed to systematically different kinds of treatments than were students who did poorly. It would be valuable if future studies could collect and analyze a larger number of student aptitude and trait measures so that this hypothesis could be examined more fully.

6. Another issue which should be considered is the fact that the analyses conducted for this report primarily examined the impact of innovative intensity on reading, language and arithmetic cognitive skills. Growth in cognitive skills was emphasized because:

- the traditional focus of national attention is on improvement in cognitive achievement, with only secondary interest in attitude changes;
- there were practical time and cost limitations and a desire to examine as intensively as possible the cognitive outcome variables; and
- the Expanded Scale Scores derived from the Comprehensive

Tests of Basic Skills were more reliable and more likely to have interval scaling properties than the study's attitude measures.

It was decided, therefore, that attitudes would be examined only in the outlier group and outlier student analyses to see if notable trends were observed in these results. Although it is recognized that educators are working toward all sorts of goals, Project LONGSTEP, like other large-scale studies of school effects, had to limit the questions that it could reasonably address and comprehensively examine. The choice was made to analyze cognitive skills.

The cognitive tests used by Project LONGSTEP, the Comprehensive Tests of Basic Skills, were designed to assess the broad concepts and abstractions developed by all curricula, as well as the basic skills required in the effective use of language and arithmetic (CTB/McGraw-Hill, 1969). Nevertheless, it is quite possible that certain of the items on these instruments may have measured outcomes that were not appropriate to the instructional goals of a particular educational program. Information was gathered concerning the instructional objectives (in arithmetic/mathematics) for a subset of the treatment groups present in the study to permit the development of a more curriculum-relevant test score or scores. Because this procedure was exploratory in nature, however, the results of the analysis of these measures are presented in a later volume.

7. One last limitation of this volume of the Project LONGSTEP final report is that the generalized impact of an innovative school environment as defined by language arts, math, social studies and science treatment approaches was not analyzed for this report. Since issues involved in these analyses are somewhat different from the more specific issues concerning the impact of language arts treatments on reading/language skills (and mathematics treatments on arithmetic skills), these school environment analyses are presented in a separate volume.

G. Overall Conclusions and Possible Implications

The single most important and well-documented finding of Project LONGSTEP was the lack of either substantial or consistent association between student achievement and overall Level of Innovation across grades. In fact, relative to the sample of schools participating in the project, students enrolled in programs with a more moderate emphasis on innovation and individualization tended to show the greatest improvement. This finding should perhaps not be viewed as surprising or regrettable. Why should change or innovation automatically mean improvement? On the contrary, it is realistic to expect--and the Project LONGSTEP data support this expectation--that some changes will be improvements, some changes will be detrimental, and some won't matter at all.

From the standpoint of time spent per day on language arts activities, it appeared that consistently overachieving students in Project LONGSTEP typically were members of programs that allocated more time to such instruction in the second grade. These findings suggest that increasing the amount of class time per day, especially in the first and second grade, may be a worthwhile strategy for improving student performance in language arts.

Our various analyses also suggest that the impact of the educational practices studied was inversely related to grade level. Specifically, it appeared that treatment-group differences, independent of the effects of student SES level, initial achievement level, and Teacher Age, SES Background and Teaching Qualifications, were greatest for second- and third-graders, less for fifth- and sixth-graders, and almost nonexistent for seventh- and eighth-graders. These findings tend to support the view that the greatest potential payoff from the investment of future educational research and instructional development funds may come from increased attention to improving education in the early elementary grades. It is suggested that further effort be expended to gain a better understanding of the educational practices that show the greatest benefit at these early grades. The results of this research could then serve as the basis for designing educational strategies that could have a relatively greater overall impact on American education.

It is less obvious why one particular aspect of much recent innovation, individualization of education, has not been shown to be clearly and consistently effective in Project LONGSTEP. Two possible explanations for this finding will be discussed, but it should be emphasized that these are purely speculative. One possibility is that not all modes of "individualization" operate effectively and that our index, Degree of Individualization, failed to tap some crucial differences between individualized programs that are effective and those that are not. This measure of individualization included four scales: Individualization in Decision Making (covering decisions in topic selection, materials selection, and sequencing decisions); Individualization of Instructional Pace; Use of Performance Agreements; and Utilization of Student Evaluation (covering the extent to which tests are used for evaluating student needs, and whether or not such evaluation serves as a basis for modifying a student's treatment).

It may be that tailoring the educational process or teaching strategy to an individual student's needs or learning style can be very effective but that allowing the student a great deal of latitude in establishing the level or standard of attainment during a school year can have a negative impact on performance (as measured by standardized achievement tests) for all but the highly motivated. The extent to which teachers monitored, encouraged, prodded, etc., their students may well be crucial but it was not tapped by our data collection instruments. Anecdotal experiences gained over the years of association with the schools participating in Project LONGSTEP suggest that some students tend to lose momentum unless the teacher provides the kind of individual attention that is unlikely except when a sufficient number of qualified assistants are available.

This problem of student motivation was discussed by Lipson (1974) in his explanation of the possible reasons why the IPI mathematics modules did not produce the dramatic gains that had been expected. To a large extent, his comments apply to a number of other individualized programs. According to Lipson, the process of each student working alone on his instructional module can have a very detrimental effect on motivation, in contrast to a conventional classroom where the group provides competition and a feeling of importance to the activities. "The student's relationship to the other members of the class and to the teacher creates a sense of obligation for some level of performance. The rhythm of the class, the rituals of class activity--even when we grumble--give our lives a pattern to follow. The progress of the class is like a tide which carries all the swimmers along even though some are slower and some are faster" (1974, p. 60). In concluding, Lipson expresses the hope that new systems will be developed "which build upon the combined strengths of the classroom approach and the modularized and individualized approaches."

The fact that our Degree of Individualization superindex failed to discriminate between successful and unsuccessful treatments also suggests the possibility that we may have overlooked some elements of individualization that occur in conventional classrooms. It could be that individualization is one of the practices that distinguishes good teachers from poor ones, even in traditional educational environments, where there are no "experimental" or "innovative" programs claiming to be individualized. Perhaps the good teacher individualizes informally and so instinctively that he or she is not even aware of it in these terms. The good teacher does not treat all children as interchangeable cogs; but rather, as the individuals they are. This kind of informal individualization would not be noted as a feature of an educational program in Project LONGSTEP--but it may nevertheless be very effective.

This hypothesis, incidentally, if correct, would extend to the elementary school level the interpretation of some loosely comparable findings at the high school level, based on Project TALENT data (Shaycoft, 1967). Analyzing Project TALENT retest data, Shaycoft found evidence, as has Project LONGSTEP, that schools differ in effectiveness. However, she failed to find, as LONGSTEP has failed to find, that these differences were meaningfully related to available variables in any way that might have been expected. She concluded:

> ... the research...suggests that schools vary in effectiveness in a great many dimensions, even though the specific school characteristics that produce these differential results are somewhat elusive. One reason

> they are so resistant to identification may be that they are elusive inherently, not just in the present context. In other words, one of the crucial differences between an effective school and an ineffective one may be something as tenuous and resistant to identification on the basis of a questionnaire as the school's atmosphere--the ambience it provides--or some other concept equally hard to nail down. Vague though this concept is, it at least has the advantage of fitting the facts and providing an explanation for them... .
>
> In summary, then,...effective high schools and ineffective ones are a fact, not a mirage. And since we have found that there are important differences among schools in their effectiveness, it would appear that education is susceptible of improvement. But... the sources of these differences among schools resist ready identification... . (Shaycoft, 1967, pp. 7-25, 7-26)

This surmise applies equally well to the LONGSTEP data.

Furthermore, we can superimpose on this hypothesis yet another level of surmise, which would go even further toward explaining the LONGSTEP findings concerning the apparent ineffectiveness of individualization and the other specific treatment variables. Let us suppose that some methods of teaching a school subject are best for some children while other methods are best for others--or possibly that some children respond best to teachers with certain personality patterns and styles of interpersonal relations while other children respond best to teachers with other personality patterns. Can one then suppose that children taught by teachers who are not "right" for them or taught by methods that do not fit their aptitudes will therefore not learn much? Not necessarily! Almost everyone who goes through the school system at least occasionally encounters poor teachers (or teachers they regard as poor) and hardly anyone has been taught by methods especially designed to be optimal for him or her. For most children it may not make a tremendous difference. Their resilience can usually overcome the effects of the occasional poor teacher or nonoptimal teaching method.

But (our hypothesis continues) this isn't true of all children. Some children have sufficiently extreme aptitude profiles (large differences between the peaks and valleys) that they may lack the flexibility to adapt well to inappropriate or nonoptimal instruction. For such children, nonoptimal instruction becomes synonymous with totally inadequate instruction.

It is possible, then, that large treatment effects were not demonstrated in Project LONGSTEP (or have not been found in many educational evaluation efforts) because such overall effects have been attenuated by aptitude-treatment interactions. The undoubtedly dramatic growth in achievement demonstrated by a number of students participating in Project LONGSTEP certainly suggests the speculation that some near-optimal match of student and treatment may have been one of the reasons for the gains of these students.

In final summary, the Project LONGSTEP findings should not be viewed as a sweeping criticism of the concepts of innovation or individualization. However, they should serve as a reminder to educators--as well as to parents and legislators--that educational innovation per se will not necessarily produce dramatic effects on student achievement. Educational quality is not synonymous with innovation or individualization. Individualization of instruction may represent a valuable approach for the improvement of American education, but the findings here suggest that individualization as a program strategy should not be viewed as the final or complete answer. In this case, more (innovation) is not necessarily better (education).

REFERENCES

Averch, H. A., Carroll, S. J., Donaldson, T. S., Kiesling, H. J., & Pincus, J. *How effective is schooling? A critical review and synthesis of research findings*. Santa Monica, California: The Rand Corp., 1972.

Benson, C. S., Schmelzle, W. K., Gustafson, R. H., & Lange, R. A. *State and local fiscal relationships in public education in California*. Report of the Senate Fact Finding Committee on Revenue and Taxation. Sacramento: Senate of the State of California, March 1965.

Boyer, E. G., Simon, A., & Karafin, G. R. *Measures of maturation*. Philadelphia: Research for Better Schools, Inc., 1973.

Bryant, E. C., Glaser, E., Hansen, M. H., & Kirsch, A. *Associations between educational outcomes and background variables*. Denver: National Assessment of Educational Progress, Education Commission of the States, 1974.

Burkhead, J., Fox, F. G., & Holland, J. W. *Input and output in large-city high schools*. Syracuse, New York: Syracuse University Press, 1967.

Campbell, D. T., & Erlebacher, A. How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In J. Hellmuth (Ed.), *Disadvantaged child, Vol. 3: Compensatory education: A national debate*. New York: Brunner/Mazel, Inc., 1970.

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), *Handbook of research on teaching*. Chicago: Rand McNally, 1963, 171-246. (Reprinted as *Experimental and quasi-experimental design for research*. Chicago: Rand McNally, 1966.)

Central Advisory Council for Education (England). *Children and their primary schools*. (Plowden Report). London: H.M. Stationery Office, 1967.

Coleman, J. S. Methods and results in the IEA studies of effects of school on learning. *Review of Educational Research*, 1975, *45*, 355-386.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. *Equality of educational opportunity*. Washington, D. C.: U. S. Government Printing Office, 1966.

Coles, G. J. Approaches to individualization: Toward a typology of innovative educational practices. Paper presented at the annual meeting of the American Educational Research Association, Washington, D. C., April 1975.

Comber, L. C., & Kunes, J. P. Science education in nineteen countries: An empirical study. New York: John Wiley & Sons, 1973.

Cronbach, L. J. Coefficient alpha and the internal structure of tests. Psychometrika, 1951, 16(3), 297-334.

CTB/McGraw-Hill. Comprehensive Tests of Basic Skills: Technical report. Monterey, California: CTB/McGraw-Hill, 1970.

CTB/McGraw-Hill. Test coordinator's handbook. Monterey, California: CTB/McGraw-Hill, 1969.

Darlington, R. B. Multiple regression in psychological research and practice. Psychological Bulletin, 1968, 69(3), 161-182.

Dyer, H. S. Some thoughts about future studies. In F. Mosteller & D. P. Moynihan, On equality of educational opportunity. New York: Vintage Books, 1972.

Echternacht, G. J. Reliability and validity of questionnaire items--status, social-psychological, and educational plans and aspirations variables. ETS Research Bulletin, 1974, RB-74-12.

Evans, J. W., & Schiller, J. How preoccupation with possible regression artifacts can lead to a faulty strategy for the evaluation of social action programs: A reply to Campbell and Erlebacher. In J. Hellmuth (Ed.), Disadvantaged child, Vol. 3: Compensatory education: A national debate. New York: Bruner/Mazel, Inc., 1970.

Farrar, D. E., & Glauber, R. R. Multicollinearity in regression analysis: The problem revisited. Review of Economics and Statistics, 1967, 49, 92-107.

Flanagan, J. C. The critical incident technique. Psychological Bulletin, 1954, 51, 327-358.

Flanagan, J. C., Dailey, J. T., Shaycoft, M. F., Gorham, W. A., Orr, D. B., & Goldberg, I. Design for a study of American youth. Boston: Houghton Mifflin, 1962.

Flanagan, J. C., Dailey, J. T., Shaycoft, M. F., Orr, D. B., & Goldberg, I. Studies of the American high school. Project TALENT monograph series, Monograph No. 2. Pittsburgh: University of Pittsburgh, 1962

Flanders, N. Analyzing teacher behavior. Reading, Mass.: Addison-Wesley, 1970.

Gardner, P. L. Scales and statistics. Review of Educational Research, 1975, 45(1), 43-57.

Glass, G. V. Teacher effectiveness. In H. J. Walberg (Ed.), Evaluating educational performance. Berkeley, Calif.: McCutchan, 1974.

Goldberger, A. S. *Econometric theory*. New York: John Wiley & Sons, 1964.

Goodlad, J. I., & Klein, M. F. *Behind the classroom door*. Worthington, Ohio: Charles A. Jones Publishing Company, 1970.

Gordon, R. A. Issues in multiple regression. *American Journal of Sociology*, 1968, *73*, 592-616.

Gullicksen, H. *Theory of mental tests*. New York: John Wiley & Sons, 1950.

Guthrie, J. W. A survey of school effectiveness studies. In U.S. Department of Health, Education, and Welfare, Office of Education, Bureau of Educational Personnel Development, *Do teachers make a difference?* Washington, D. C.: U. S. Government Printing Office, 1970.

Hammond, J. L. Two sources of error in ecological correlations. *American Sociological Review*, 1973, *38*, 764-777.

Hanushek, E. A. The production of education, teacher quality, and efficiency. In U.S. Department of Health, Education, and Welfare, Office of Education, Bureau of Educational Personnel Development, *Do teachers make a difference?* Washington, D. C.: U. S. Government Prining Office, 1970.

Hawkridge, D. G., Chalupsky, A. B., & Roberts, A. O. H. *A study of selected exemplary programs for the education of disadvantaged children*. Palo Alto, California: American Institutes for Research, 1968.

Hertzberg, A., & Stone, E. *Schools are for children: An American approach to the open classroom*. New York: Schocken Books, 1971.

Hilton, T. L., Beaton, A. E. & Bower, C. P. *Stability and instability in academic growth--a compilation of longitudinal data*. Princton: Educational Testing Service, 1971.

Husen, T. (Ed.). *International study of achievement in mathematics*. New York: John Wiley & Sons, 1967.

Jencks, C., Smith, M., Acland, H., Bane, M. J., Cohen, D., Gintis, H., Heyns, B., & Michelson, S. *Inequality, A reassessment of the effect of family and schooling in America*. New York: Basic Books, 1972.

Kelley, T. L. *Fundamentals of statistics*. Cambridge, Mass.: Harvard University Press, 1947.

Kerlinger, F. N., & Pedhazur, E. J. *Multiple regression in behavioral research*. New York: John Wiley & Sons, 1973.

Kiesling, H. J. Measuring a local government service: A study of school districts in New York State. *The Review of Economics and Statistics*, 1967, *49*, 356-367.

Klitgaard, R. E., & Hall, G. R. *A statistical search for unusually effective schools*. Santa Monica, California: The Rand Corp., 1973.

Lee, A. M., & Holley, F. M. An ideal evaluation design in a public school setting: Or, where are you,Campbell and Stanley,now that we need you? Paper presented at the meeting of the American Educational Research Association, Washington, D. C., April 1975.

Linn, R. L., Werts, C. E., & Tucker, L. R. The interpretation of regression coefficients in a school effects model. *Educational and Psychological Measurement*, 1971, *31*, 85-93.

Lipson, J. I. IPI math--An example of what's right and wrong with individualized modular programs. *Learning*, March 1974, *2*(7), 60-61.

Lord, F. M. A paradox in the interpretation of group comparisons. *Psychological Bulletin*, 1967, *68*, 304-305.

Lord, F. M., & Novick, M. R. *Statistical theories of mental test scores*. Reading, Mass.: Addison-Wesley, 1968.

Mayeske, G. W., Okada, T., & Beaton, A. E., Jr. *A study of the attitude toward life of our nation's students*. Washington, D. C.: U. S. Government Printing Office, 1973.

Mayeske, G. W., Okada, T., Beaton, A. E., Jr., Cohen, W. M., & Wisler, C. E. *A study of the achievement of our nation's students*. Washington, D.C.: U. S. Government Printing Office, 1973.

Mayeske, G. W., Wisler, C. E., Beaton, A. E., Jr., Weinfeld, F. D., Cohen, W. M., Okada, T., Proshek, J. M., & Tabler, K. A. *A study of our nation's schools*. Washington, D. C.: U. S. Government Printing Office, 1972.

Michigan State Department of Education. Research into the correlates of school performance: A review and summary of literature. Paper presented at the annual meeting of the American Educational Research Association, New York, 1971.

Mollenkopf, W. G., & Melville, S. D. A study of secondary school characteristics as related to test scores. *Research Bulletin 56-6*. Princeton: Educational Testing Service, 1956. (mimeograph)

Mood, A. M. Partitioning variance in multiple regression analyses as a tool for developing learning models. *American Educational Research Journal*, 1971, *8*(2), 191-202.

Mosteller, R., & Moynihan, D. P. (Eds.). *On equality of educational opportunity*. New York: Random House, 1972.

O'Connor, E. F. Unraveling Lord's paradox: The appropriate use of multiple regression analysis in quasi-experimental research. Research Bulletin (No. RB-73-53). Princeton: Educational Testing Service, August 1973. (a)

O'Connor, E. F. Why matching, analysis of covariance, and multiple regression analysis generally fail to control for preexisting differences. Paper presented at the annual meeting of the Western Psychological Association, Anaheim, California, 1973. (b)

O'Connor, E. F., & Klein, S. P. A critique of the statistical analysis of the OEO experiment in performance contracting. Research Bulletin (No. RB-73-61). Princeton: Educational Testing Service, October 1973.

Peaker, G. F. The Plowden children four years later. London: National Foundation for Educational Research in England and Wales, 1971.

Pohlman, J. T., & Newman, I. Regression effects when the assumption of rectilinearity is not tenable. Multiple Linear Regression Viewpoints, 1973, 4, 51-59.

Purves, A. C. Literature education in ten countries. New York: John Wiley & Sons, 1973.

Ribich, T. I. Education and poverty. Washington, D. C.: Brookings Institution, 1968.

RMC, Inc. Evaluation of the Emergency School Assistance Program, Final Report UR-163, Volume II: Design and Findings of Phase I Survey. Bethesda, Maryland: Author, 1971. (Contract No. OEC-0-71-0845)

Robinson, W. S. Ecological correlations and the behavior of individuals. American Sociological Review, 1950, 40, 351-357.

Rosenshine, B. The stability of teacher effects upon student achievement. Review of Educational Research, 1970, 40(5), 647-662.

Rubin, D., Trismen, D. A., Wilder, G., & Yates, A. A descriptive and analytic study of compensatory reading programs. Princeton, N. J.: Educational Testing Service, 1973.

Schramm, W. (Ed.) Four case studies of programed instruction. New York: Fund for the Advancement of Education, June 1964.

Shaycoft, M. F. The high school years: Growth in cognitive skills. (Interim report 3 to the U. S. Office of Education, Cooperative Research Project No. 3051). Pittsburgh: American Institutes for Research and University of Pittsburgh, 1967.

Simon, A., & Boyer, E. G. (Eds.). Mirrors for behavior. Philadelphia: Research for Better Schools, Inc., 1967.

Snow, R. E. Representative and quasi-representative designs for research on teaching. Review of Educational Research, 1974, 44, 265-291.

Spådek, B. Open education. New York: National Association for the Education of Young Children, 1970.

Spady, W. G. The impact of school resources. In F. N. Kerlinger (Ed.), Review of research in education, No. 1. Itasca, Illinois: Peacock Publishers, 1973.

Stallings, J. An implementation study of seven FOLLOW THROUGH models for education. Paper presented at the American Educational Research Association Annual Meeting, Chicago, 1974.

Tatsuoka, M. M. Multivariate analysis in educational research. In F. N. Kerlinger (Ed.), Review of research in education, No. 1. Itasca, Illinois: Peacock Publishers, 1973.

Thomas, J. A. Efficiency in education: A study of the relationship between selected inputs and mean test scores in a sample of senior high schools. Unpublished Ph.D. dissertation, Stanford University, School of Education, 1962.

U. S. Office of Education. Pacesetters in innovation. Washington, D. C.: U. S. Government Printing Office, 1969.

Walberg, H. J., & Thomas, S. C. Characteristics of open education: Toward an operational definition. Newton, Mass.: TDR Associates, Inc., 1971.

Weisgerber, R. A. (Ed.). Developmental efforts in individualized learning. Itasca, Illinois: Peacock Publishers, 1971. (a)

Weisgerber, R. A. (Ed.). Perspectives in individualized learning. Itasca, Illinois: Peacock Publishers, 1971. (b)

Werts, C. E., & Linn, R. L. A general linear model for studying growth. Psychological Bulletin, 1970, 73(1), 17-22.

Wiley, D. E. Design and analysis of evaluation studies. In M. C. Wittrock and D. E. Wiley (Eds.), The evaluation of instruction: Issues and problems. New York: Holt, Rinehart and Winston, 1970, 259-288.

Wiley, D. E. Another hour, another day; Quantity of schooling, a potent path for policy. Chicago: University of Chicago, 1973. (mimeograph)

Wiley, D. E., & Bock, R. D. Quasi-experimentation in educational settings: Comment. The School Review, 1967, 75, 353-366.

Womer, F. B. What is national assessment? Denver: National Assessment of Educational Progress, 1970.